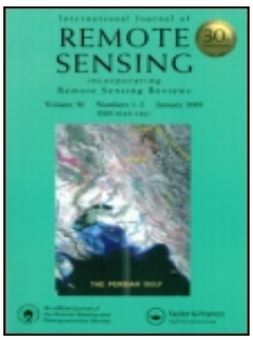

International Journal of Remote Sensing

ISSN: 0143-1161 (Print) 1366-5901 (Online) Journal homepage: https://www.tandfonline.com/loi/tres20

Quality assessment of image classification algorithms for land-cover mapping: A review and a proposal for a cost-based approach

P. C. Smits , S. G. Dellepiane & R. A. Schowengerdt

To cite this article: P. C. Smits , S. G. Dellepiane & R. A. Schowengerdt (1999) Quality assessment of image classification algorithms for land-cover mapping: A review and a proposal for a cost-based approach, International Journal of Remote Sensing, 20:8, 1461-1486, DOI: 10.1080/014311699212560

To link to this article: https://doi.org/10.1080/014311699212560

INT. J. REMOTE SENSING, 1999, VOL. 20, NO. 8, 1461–1486

Quality assessment of image classification algorithms for land-cover mapping: a review and a proposal for a cost-based approach

P. C. SMITS, S. G. DELLEPIANE

Department of Biophysical and Electronic Engineering, University of Genoa Via all'Opera Pia 11A, 16145 Genoa, Italy; e-mail: smits@dibe.unige.it

and R. A. SCHOWENGERDT

Department of Electrical and Computer Engineering, University of Arizona, Tucson AZ 85721, USA

(*Received 23 June 1997; in final form 13 July*)

Abstract. The following issues relate to quality assessment of image classification: the classification methods as such, the methods to evaluate the classification results, and the requirements of the application. In this paper, a number of evaluation methods are reviewed, and it is concluded that those based on confusion matrices and the KHAT analysis are the most suited if one is interested in comparing classifiers. The novelty of this paper is that much attention is given to the subjectivity present in every evaluation scheme, and that the concept of accuracy is extended to quality by creating the link between accuracy, objectives, and costs. A protocol is proposed for quality assessment related to the economical reality. An example based on a hypothetical data set shows that the economic cost of misclassification can be high, and that it may be advantageous for the user to reconsider either the objectives, the type of data used, or other aspects of the remote-sensing system that he uses to produce the map.

1. Introduction

This paper addresses the issue of *quality* assessment of remote-sensing image classification systems for land-cover mapping. Maling (1989), in his chapter on the concept of the accurate map, rightly recognizes that maps are deliberate generalizations of reality, and that all survey and map production processes inevitably introduce errors. The user of land-cover maps needs to know how *accurate* the product is in order to use the data efficiently. Accuracy describes the closeness of a measurement to the true value of the quantity being measured. With the term quality, we put the concept of accuracy in a broader context, by attaching to it a label which refers directly to the objectives and requirements that have been defined prior to the accuracy assessment. A complete quality evaluation system should therefore be able to tell end-users (i.e. those who will actually use the *thematic maps* created with the image processing tools) the actual cost of choosing a certain data set, feature vector, or classification algorithm.

Although a large amount of literature on accuracy assessment is available from the earth science, engineering and statistics communities, it is felt that in the

International Journal of Remote Sensing
ISSN 0143-1161 print/ISSN 1366-5901 online
http://www.tandf.co.uk/JNLS/res.htm
http://www.taylorandfrancis.com/JNLS/res.htm

remote-sensing community, especially among users of Geographical Information Systems (GIS) land-cover products (related to the physical properties) and land-use products (related to human activities), there is a need for awareness of both accuracy and quality of image classification algorithms. The present lack of accuracy-awareness has two implications. The first, which is consequential in the short term for the end-user himself, is an inefficient use of expensive remote-sensing *data*. The inefficiency finds its origin, for example, in the classification algorithms not being suited to certain tasks, thus causing unnecessary errors of omission and commission. The second effect is of a more long term kind, and is hazardous for the technology providers: the classification *algorithms* are utilized improperly. New types of data call for appropriate classification algorithms. A good example is the recent availability of relatively inexpensive synthetic aperture radar (SAR) data from space borne platforms for civil applications. Traditional classifiers, such as K-nearest neighbour (KNN) or maximum likelihood (ML), may perform well on Landsat-TM data sets, but may not be appropriate for the backscatter radar signals of SAR with its speckle noise and different remote-sensing phenomenology. A lack of awareness regarding real accuracy may cause the user to apply familiar classifiers because for instance they are supported by the commercial GIS software, rather than less well known but more effective algorithms.

One solution to the above problems is to provide the end-users with appropriate information and tools to evaluate classifiers. But the responsibility lies also in the hands of authors presenting new methods. Although the importance of thoroughly evaluating machine vision systems has been often stressed (Jain and Binford 1991, Haralick and Shapiro 1993, Kanungo *et al.* 1995), remote-sensing image classification results are often poorly founded, and not compared with those of traditional methods. Examples of limited test data, or training data used as test data, are not difficult to find although various textbooks have long stressed the importance of a proper evaluation (see, for instance, Swain 1978, Mather 1987, and Jensen 1996).

Sometimes the difference between image classification and segmentation is very small as happens in cases where the segmentation algorithms label pixels with a semantic image class label. Jain and Binford (1991) describe the underlying reason for the poor acceptance of the field of image segmentation as characterized by the fact that, although a number of competing and promising approaches to image segmentation exist, a thorough comparative evaluation and critical assessment regarding issues such as generality, theoretical foundation, stability with respect to intrinsic parameter variations and input image quality had not been achieved on a global scale. Similar considerations would also hold for the subjects of classification (Duin 1996), and image processing in general (Zamperoni 1996), and therefore also for remote-sensing image classification. Clearly, a strong involvement of the end-user is a *conditio sine qua non*. Some remote-sensing articles on accuracy assessment do involve the end-user (Ginevan 1979, Aronoff 1982a, Story and Congalton 1986), but often merely address the issue of how end-users would like to have the results presented when assessing the accuracy of a classifier, rather than explicitly dealing with what the end-user actually needs.

In order to increase the awareness among end-users, it is important that they realize the actual costs if decisions are based on erroneous information. This brings the total number of issues related to the subject of quality assessment to three: the classification methods as such, the methods to evaluate these algorithms, and the relation between the accuracies of the results and the objectives and requirements

formulated by the end-users; i.e. those who will benefit from the classification results. These issues form the structure of the present paper.

Before proceeding, it is appropriate to make some general remarks. The work of Lunetta *et al.* (1991) gives a general view of the errors associated with the various processes typical of a remote-sensing-based decision support system (figure 1). Rather than reviewing accuracy assessment methods, Lunetta *et al.* suggest priorities for error quantification research topics, including the development of standardized and more cost-effective remote-sensing accuracy assessment procedures, development of ground data collection guidelines, procedures for vector-to-raster and raster-to-vector conversions, assessment of scaling issues for the incorporation of elevation data in georeferencing, and development of standardized geometric and thematic reliability legends.

The remote-sensing community seems ready for standardization regarding assessment and reporting of accuracy, and to address the issues regarding error and accuracy in GIS and related aspects, including remote-sensing image classification (for recent articles see, for instance, Congalton 1991, Lunetta *et al.* 1991, Janssen and Van der Wel 1994, Estes and Mooneyhan 1994, Gopal and Woodcock 1994, Ma and Redmond 1995a,b).

In the recent literature on accuracy assessment, some convergence is observed regarding standardization. Most current research employs *confusion matrices* (also called error matrixes, confusion tables, and contingency matrices). Once an accurate confusion matrix is produced, various methods and coefficients are available to analyse it (Congalton 1991, Ma and Redmond 1995a, Zhuang *et al.* 1995, Richards 1996) (see figure 2). However, obtaining a reliable confusion matrix remains a relatively weak link in the whole chain of accuracy assessment. Potential problems are the subjectivity inevitably induced by the choice of the classification scheme (labels), the training samples (in the case of supervised classification), and the reference data sampling size and strategy.

When the classes are well separated in a feature space and there is no overlap between the distributions of the categories, most classifiers should return the same result, which will hardly depend on the choice of training sample. When the classes are confused, which is often the case in a real world data set corrupted by different types of noise (Landgrebe and Malaret 1986), the classifiers may disagree, depending on the *a priori* information that is incorporated into the classification models. Moreover, even a given classifier can produce different results when trained with a different data set. At this stage it becomes most interesting to see which classifier performs best with the given data set.

We propose in this paper a solution for integrating the subjective and objective

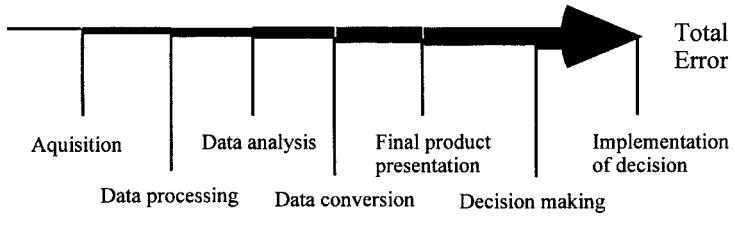

Figure 1. The accumulation of error in a typical remote-sensing information processing flow (after Lunetta *et al.* 1991).

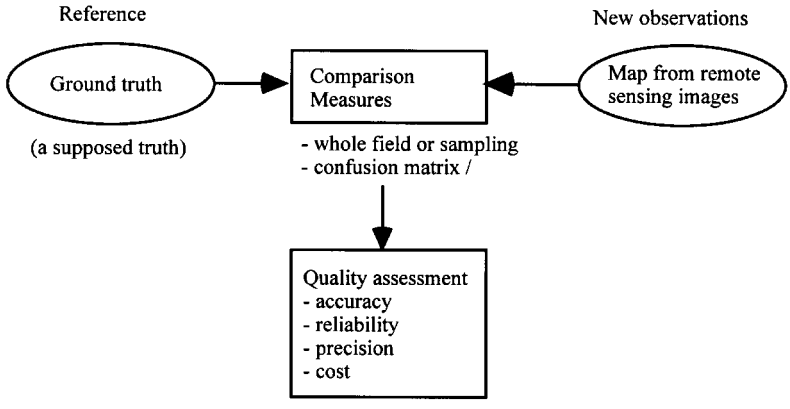

Figure 2. From reference and observed data to quality assessment.

components necessary for a valid accuracy assessment. The notion of economic costs, directly derived from requirements and needs of end-users, are related to the assessed accuracy. The objectives of this work are first, to present an overview of the classification accuracy assessment methods that are available, and second, to relate the requirements of the end-user of the classification product to the concept of accuracy. The focus of attention will be on the issues related to evaluation and accuracy from the end-user's perspective, thus creating a system for quality assessment.

One benefit of such an analysis is a better understanding of *who* the actual end-user is, and *what* his needs and requirements are (figure 3). This is important to also moderate the skepticism towards digital data and tools that still lingers among workers in the field (Maling 1989).

In the literature, many (remote-sensing) image classification techniques have been proposed, although commercially only a few approaches are frequently found (see table 1). We limit ourselves to some relevant considerations related to the difficulties in image classification in §2. Section 3 reviews methods for the accuracy assessment of segmentation and classification methods, in which the normalized confusion matrix has a prominent role. Section 4 connects the normalized confusion matrix to a *cost matrix* that provides a cost estimate for the application of a certain classification algorithm; moreover, using an example (§5), it is suggested how prospective users of

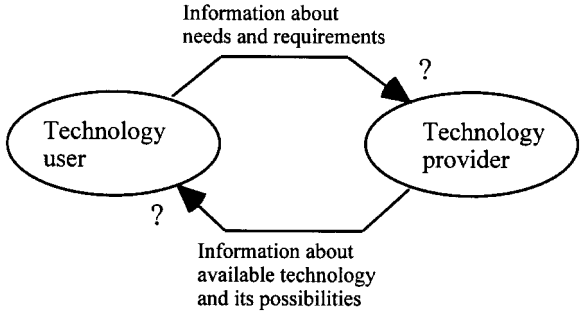

Figure 3. Mutual understanding between technology users and providers is needed for an efficient, goal-directed development of technology.

Table 1. Frequently found classification algorithms in commercial GISs.

Classifier	Advantages	Disadvantages
Parellelipiped	Fast and simple; distribution free	Due to corners, pixels may be classified from spectral mean
Minimum distance	No unclassed pixels; fast	Does not consider class covariance
Mahalanobis	Takes the class covariance into account	Over classification of signatures with large values in covariance matrix; parametric: assumes normal distribution
ML/Bayesian	Relatively accurate; variability of classes is taken into account	Computationally expensive; assumes normal distribution; overclassification of signatures with large values in cov. matrix
Neural network	Distribution-free; fast after training	Slow training; no theoretical basis; stochastic convergence

classifiers and maps could achieve a systematic assessment of such a matrix. The paper concludes with a summary in §6.

2. Difficulties in classification of remote-sensing images

Pattern identification is the process in which a decision rule is applied (Haralick and Shapiro 1993). Basically, a pattern identification process produces a category identification sequence that assigns to every object a category, based on the data or observations. This pattern identification is also called pattern classification or simply classification. In remote-sensing image analysis, classification uses suitable algorithms to label the pixels in an image as representing particular ground cover types, or classes (Richards 1993).

Many difficulties originate from the observed scene, which in general does not consist of neatly spaced, homogeneous parcels of land, nor are man-made structures built to fit the raster of some sensor, implicating corruption of the class signatures. In some cases the effect of these phenomena can be modelled and reduced using linear mixture models, class probability, or fuzzy membership functions (Foody and Cox 1994). As observed earlier, however, some level of generalization is necessary, and for the development of this paper it is assumed that each pixel belongs to one class only, accepting a certain degree of error, and realizing that this may have implications for the applicability of very coarse data (e.g. Advanced Very High Resolution Radiometer (AVHRR) data).

The selected features should be chosen by optimizing a criterion, estimated from the training data. How well these training data represent a class is an important question: do they adequately sample the feature space for each class? Moreover, to fully use the information contained in the feature measurements, training samples are needed from all the classes of interest.

The problem of labeling inconsistencies regards the *representativeness* of the training data and is due to mixed pixels (class overlap), transition zones, dynamic zones, within-class variability (covariance), limited training data, and topographic shading, just to name a few. This type of error is difficult to quantify. The other type of error, the *classification-induced* error, can be reduced using carefully defined classes and number of classes, classification schemes, and the choice of the feature vector. For example, there exists a strong relationship between the number of classes, the optimum number of features and the size of the training set. In pattern recognition

it is well known that when the ratio of the number of training samples to the number of feature measurements is small, the estimates of the discriminant functions are not accurate, which may influence the quality of the result. This is the so-called Hughes phenomenon, which is discussed for hyperspectral data in (Shahshahani and Landgrebe 1994).

3. Accuracy assessment of classifiers for remote-sensing imagery

In the literature, the term accuracy is commonly utilized as if one could objectively and beyond any doubt determine the relation between a so-called data signature in the imagery and a category. This is valid only in a few cases, where the physical parameters used to describe the training and the test sites correspond to the physical parameters sensed by the remote-sensing instrument (figure 4). In most cases this relation is, to say the least, not guaranteed. A clear example of this is the use of the radar back-scatter signal of SAR instruments for classification purposes. The angle of incidence and thus the flight direction of the platform has great influence on the back-scattered signal. In such a case, it would be more appropriate to talk in terms of *map consistency*, or even *signature consistency*, rather than accuracy. In an application like remote-sensing, the accuracy of a classifier in reality is a *relative accuracy*.

Although some authors suggest the use of the probabilistic or fuzzy outcome of a classifier in terms of, for instance, information score (Kononenko and Bratko 1991), in this article we will focus on classifiers that produce a hard decision, i.e. where labelling is mutually exclusive and each pixel is assigned only one label.

3.1. *Premises for accuracy assessment*

The fundamental aspects of any evaluation scheme are (*a*) the ground truth data, (*b*) the classification scheme and category semantics, and (*c*) the sample scheme and size; the following subsections address these aspects.

(*a*) *Ground truth data*. The collection of ground truth data is necessary for both the training and validation, and should be seen as an integral part of the accuracy assessment of an image classification method. For an adequate calculation of the classification accuracy, a high quality test set, different from the training set, should

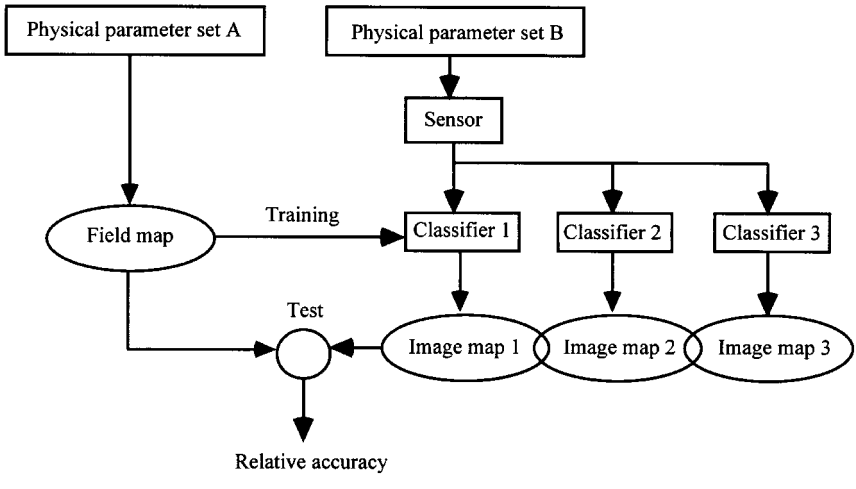

Figure 4. What does accuracy really mean?

be used. Users should be cautious in interpreting the results if the ground data from which the test pixels were identified were not collected on the same date as the remotely sensed image.

A possible strategy could be to use as reference data information that comes from a source that is presumed to be accurate enough, for instance, photo-interpretation or aerial reconnaissance (Congalton 1991). It is clear that this strategy may underestimate (or overestimate) the true accuracy of the classification system because of unknown inaccuracies in the reference data. This refers to the earlier mentioned problem of the reference data being derived from a physically different source than the images (figure 2).

(*b*) *Classification scheme and category semantics.* Congalton (1991) gives a number of simple guidelines for setting-up a classification scheme. First, any area to be classified should fall into one and only one category, and every area should be included in the classification, and second, the classification scheme should be hierarchical in nature, such that certain categories within the classification scheme can be collapsed to form more general categories. The scheme should therefore be logical and easy to understand.

In certain cases it may be possible and more advantageous to not use the semantics of class labels, such as those used by the United States Geological Survey (USGS), but to replace the label values by the spectral mean values of the original imagery. This technique proposed in the data compression literature (Lee and Chen 1995), allows one to reduce drastically the information while preserving the original spectral meaning of the data, without having to generalize or reduce the spectral mixture of an area to one label.

(*c*) *Sample scheme and size.* The sampling scheme is used to obtain an adequate reference data set and selects inside an area those points that serve as reference data. Important requisites of a sample scheme are that it sufficiently represents every category on the land-use or land-cover map, and that no biases will be introduced by the sample scheme itself. Table 2 gives a summary of sample strategies proposed in the literature.

Fitzpatrick-Lins (1981) addressed the issue of the sampling procedure needed to assess the accuracy of land-use and land-cover maps. She compared a manual stratified, systematic unaligned (SSU) sample technique with a computer-based one, followed by a random selection stratified by category. In order to estimate the approximate sample size needed, the normal approximation equation was used. Congalton (1991) observed that, although the sample size found by this technique can be appropriate for computing overall accuracy, it is not an appropriate for confusion matrices.

Simple random sampling is not suitable to represent relatively small categories, which suffer undersampling. Stratified random sampling that uses *a priori* knowledge of the true map category marginal proportions may overcome this problem (Card 1982). In this procedure, the test data is collected after the classification has been done, such that one can collect the appropriate number of samples per category.

A drawback of any random sampling strategy is that the collection of ground information can be tedious and expensive, because the ground points may not be easy to locate. In many cases, it is necessary to remain close to roads, which biases the sampling (Thomas and Allcock 1984).

Summarizing this section, it is clear that, although many methods have been proposed to eliminate the subjectivity right from the beginning by carefully

Table 2. Sampling strategies proposed (√) to evaluate the qualitative accuracy of Landsat-derived maps (modified after Maling 1989).

	Sampling strategy				
	Unrestricted		Stratified		
Author	Random	Systematic	Random	Unaligned systematic	Combined/ multistage
Hord and Brooner (1976)	√	–	–	–	–
Lins (1978)	–	√	–	–	–
Fitzpatrick (1977)	–	–	√	–	–
Van Genderen *et al.* (1978)	–	–	√	–	–
Latham (1979)	–	√	–	–	–
Hay (1979)	√	–	–	–	√(1)
Ginevan (1979)	√(3)	–	–	–	–
Rosenfield and Melley (1980)	–	–	–	√	–
Strahler (1981)	–	–	√	–	–
Fitzpatrick-Lins (1980, 1981)	–	–	–	√	–
Aronoff (1982a)	√(3)	–	–	–	–
Aronoff (1982b)	√(3)	–	–	–	√(1)
Card (1982)	√		√		
Rosenfield *et al.* (1982)	–	–	–	–	√(2)
Stehman (1996)	–	–	√	–	–

(1) Simple unrestricted random sampling followed by stratification by class and additional random sampling within each category until the minimum sample size has been attained.

(2) Stratified unaligned systematic sampling followed by additional random sampling in the under-represented categories.

(3) Subsequent analyses are based upon acceptance sampling methods.

addressing all kinds of uncertainty factors, for practical reasons it will be impossible to succeed completely. Therefore, in the next section methods that will eliminate as far as possible the subjectivity induced into the confusion matrix will get extra attention.

3.2. *Evaluation methods*

Table 3 gives an overview of the accuracy assessment methods discussed in this section. The next subsections are organized according to the category to which a method belongs: (*a*) based on confusion matrices, (*b*) fuzzy techniques, (*c*) receiver operating characteristics, and (*d*) other techniques.

(*a*) *Confusion matrix based techniques*. The importance of confusion matrices is stressed in Congalton (1991). A confusion matrix is a square array of numbers set out in rows and columns which express the number of sample units assigned to a particular category relative to the actual category as verified by ground truth information or reference data set. Basically, a confusion matrix allows both *descriptive* and *analytical* analysis.

Descriptive techniques are relatively simple and include computation of the overall accuracy (division of the total correct by the total number of units) and the individual class accuracy. The latter measure can be expressed in two ways: by calculating the *producer's accuracy*, which is based on the reference data (error of omission), or the *user's accuracy*, based on the total number of pixels classified in specific classes (commission error).

Table 3. An overview of the discussed classification assessment methods.

Authors	Key words	Overall accuracy	Class accuracy	Classifier comparison
Rosenfield (1981)	Variance techniques	√	–	–
Aronoff (1982a)	Producer's risk and user's risk	√	√	–
Maxim and Harrington (1983)	Pseudo-Bayesian estimates in multi-variate analysis	√	–	–
Congalton (1991)	Confusion matrices	√	√	–
Czaplewski (1992), Zhu *et al.* (1996)	Multi-variate composite estimator	√	–	–
Gopal and Woodcock (1994), Woodcock (1996)	Fuzzy analysis	√	√	√
Kanungo *et al.* (1995)	Operating characteristics	–	√	–
Ma and Redmond (1995a)	tau coefficient	√	√	√
Zhuang *et al.* (1995)	Tukey analysis			√
Richards (1996)	Bayesian estimates of map accuracy	√	√	–
This paper	Confusion matrices, RS system evaluation estimated absolute cost of error	√	√	√

Analytical statistical techniques are useful for comparing different classification methods. One type involves discrete multi-variate techniques for statistical tests on the classification accuracy (Congalton *et al.* 1983).

Normalizing the confusion matrix by iterative proportional fitting which forces each row and column to sum to one (Bishop *et al.* 1975), allows one to compare directly both the main-diagonal and the off-diagonal values of confusion matrices obtained by different classification algorithms (or by different analysts). *The normalized overall accuracy*, obtained in a similar way as the overall accuracy which uses the raw confusion matrices, also contains information about the off-diagonal cell values and may therefore be a more accurate representation.

Kappa analysis (Cohen 1960, Congalton and Mead 1983, Stehman 1996) is a currently popular multi-variate technique for accuracy assessment. The estimate of kappa is the so-called *KHAT statistic*, and gives a measure that indicates if the confusion matrix is significantly different from a random result. The kappa analysis also can be used to compare different matrices from different classifiers and to determine if one result is significantly better than the other. In case of insignificant difference, other aspects can be taken into account, such as computational load of the method (Kononenko and Bratko 1991).

The KHAT statistic is computed as (Congalton 1991)

$$\hat{K} = \frac{N \sum_{i=1}^{r} x_{ii} - \sum_{i=1}^{r} (x_{i+} x_{+i})}{N^2 - \sum_{i=1}^{r} (x_{i+} x_{+i})} \tag{1}$$

where r is the number of rows in the matrix, x_{ii} is the number of observations in row i and column i, x_{i+} and x_{+i} are the marginal totals of row i and column i,

respectively, and N is the total number of observations. The KHAT statistic assumes a multinomial sampling model and that the variance is derived using the Delta model. Confidence intervals can be calculated for KHAT using the approximate large sample variance:

$$\hat{\sigma}(\hat{K}) = \frac{1}{N}\left[\frac{\theta_1(1-\theta_1)}{(1-\theta_1)^2} + \frac{2(1-\theta_1)(2\theta_1\theta_2-\theta_3)}{(1-\theta_2)^3} + \frac{(1-\theta_1)^2(\theta_4-4\theta_2)^2}{(1-\theta_2)^4}\right] \quad (2)$$

where $\theta_1 = \sum_{i=1} \frac{x_{ii}}{N}$, $\theta_2 = \sum_{i=1} \frac{x_{i+}.x_{+i}}{N^2}$, $\theta_3 = \sum_{i=1} x_{ii}(x_{i+}+x_{+i})/N^2$, and

$\theta_4 = \sum_{i=1,j=1} x_{ij}(x_{j+}+x_{+i})^2/N^3$.

Two results can be compared by using a test for significant difference. For large samples, this so-called *Z statistic* test is given by

$$Z \approx \frac{\hat{K}_1 - \hat{K}_2}{\sqrt{\hat{\sigma}_1^2 + \hat{\sigma}_2^2}}. \quad (3)$$

Stehman (1996) presents formulae for estimating the kappa coefficient and its variance for stratified random sampling. Briefly, given a census of all the N pixels of a remote-sensing image classified into q classes, and the true classification of each pixel, the *population* confusion matrix is

		Reference				
		1	2	...	q	Row total
	1	N_{11}	N_{12}	...	N_{1q}	N_1
Image	2	N_{21}	N_{22}	...	N_{2q}	N_2
(Stratum)	:	:	:	...	:	:
	q	N_{q1}	N_{q2}	...	N_{qq}	N_q
Column total		M_1	M_2	...	M_q	N

Since checking each classified pixel is not only impractical but also undesirable, one must use sampling to estimate kappa, and the above table would be replaced by the *sample* confusion matrix. Now, the parameter of interest, K, can be estimated as

$$\hat{K} = \frac{N\sum_{h=1}^{q} \hat{N}_{hh} - \sum_{h=1}^{q} N_h \hat{M}_h}{N^2 - \sum_{h=1}^{q} N_h \hat{M}_h} \quad (4)$$

where $\hat{N}_{hh} = \frac{N_h}{n_h} n_{hh}$ is an unbiased estimator of N_{hh}. For any column $j = 1, ..., q$, an unbiased estimator of M_j is $\hat{M}_j = \sum_{h=1}^{q} \frac{N_h}{n_h} n_{hj}$. Stehman provides the estimated variance of the kappa estimator $\hat{K}$, and shows in a detailed empirical study that these estimators have little bias, and confidence intervals perform well, even at relatively small sample sizes. For more details on the kappa analysis, the reader is referred to (Cohen 1960, Hudson and Ramm 1987, Congalton 1991, Shehman 1996).

The actual implementation of accuracy assessment methods may cause practical problems. The accepted conventional method for accuracy assessment, as reviewed in Congalton (1991), requires that the reference data and the classification result share a common classification scheme. Zhu *et al.* (1996) argue that this method may be well-suited to relatively fine resolutions, but not for coarse resolutions, such as produced by the AVHRR. The alternative they propose for this type of data is the *multi-variate composite estimator* (Czaplewski 1992). The approach accommodates the situation where different classification schemes are used by the remote-sensing product and the reference data, and it allows for more complex sampling designs beyond the simple random sample of point plots required by the conventional method.

Ma and Redmond (1995a) present the *tau coefficient* to the remote-sensing community. This coefficient looks promising since it seems to better represent the percentage of agreement than does kappa, and is easier to calculate and interpret. The main difference between the kappa and the tau coefficients is that KAPPA is based on the *a posteriori* probabilities of group membership, whereas tau is based on the *a priori* probabilities. Two distinct tau coefficients can be calculated: one for use with classifications based on equal probabilities of group membership, and one based on unequal probabilities of group membership. Ma and Redmond give three justifications for the use of the tau coefficients:

1. Conceptually, tau is easier to understand and interpret than kappa.
2. Both tau and its variance estimates are relatively simple to calculate.
3. Unlike kappa, tau compensates for unequal probabilities of groups or for different number of groups, depending on the version of tau utilized (equal or unequal probabilities of group membership).

The disadvantage of the tau method is that it makes use of *a priori* probabilities, which have to be estimated, or computed, from the available data, which gives a different look on the advantages.

Another statistical confusion matrix-based method is reported by Zhuang *et al.* (1995). They introduce the *Tukey multiple comparison method* and argue that multiple comparison of results from the new classifiers with those from conventional classifiers is needed.

The Tukey multiple comparison method is based on a distance measure of performance. It uses the main diagonals of the normalized confusion matrices (one for each classifier in the test) in a distance-like measure, which produces for each classifier a relative distance to the 'mean' performance. Then, multiple comparisons can be made using the Tukey critical distance, which depends on the critical value of the Studentized range at a given risk level, the number of classifiers, the mean-squared error (MSE), the number of degrees of freedom associated with MSE, and the number of observations (i.e. categories) of the classifiers. The paper gives rather detailed explanations on the actual implementation of the proportional fitting procedure (Fienberg 1970), which includes the elimination of zero counts (Fienberg and Holland 1970).

Zhuang *et al.* conclude that the major advantage is that the comparisons can be done in one shot. Although the most interesting possibility of applying the method for pairwise comparisons is stressed as a key-feature, the paper does not provide an example.

Richards (1996) makes a clear and explicit distinction between testing a classifier

and testing a thematic map. The motivation is that the difference between user's accuracy and producer's accuracy, as proposed by Story and Congalton (1986), does not sufficiently account for the true class proportions on the ground (represented by the so-called prior probabilities of classes). Congalton's user's accuracy is only applicable as a map accuracy indicator, therefore, if (*i*) the compositions of a test set of pixels, used to assess classifier performance, reflects true class proportions, or (*ii*) samples are taken from the map itself, they then represent a true indicator of class proportions in the map.

In order to address this issue, Richards applies Bayes' theorem to obtain the probability that the class Y indicated on the map is really representing the true class X:

$$p(x = X/y = Y) = \frac{p(y = Y|x = X)p(x = X)}{p(y = Y)}, \tag{5}$$

where $p(x = X)$ is the probability of choosing a pixel from class X at random from the imaged region. $p(y = Y)$ is the probability that class Y occurs in the thematic map and can be computed from

$$p(y = Y) = \sum_{Z \in \{A, B, C\}} p(y = Y|x = Z)p(x = Z), \tag{6}$$

$\{A, B, C\}$ being the set of possible classes; $p(y = Y|x = X)$ represents Congalton's producer's accuracy.

However, since generally the testing set of pixels does not reflect the class prior probabilities, the map accuracy either needs to be determined from the performance of the classifier, or could be approximated by the class proportions in the thematic map, which is acceptable if the classifier performs well on all classes (i.e. $p(x = Z) = p(y = Z)$). The above demonstrates that it is difficult to obtain reliable estimates for $p(x = X/y = Y)$ using testing set data in the classical sense.

Another type of approach comes from the information-theory community, and uses measures of entropy and information, frequently used in digital communications. Kew (1996) applies this concept to the assessment of remote-sensing image classification, formulated in terms of a noisy information channel. Using the terms mutual information and relative information, the available ground truth is related to the classifier output. Although the method seems to offer interesting properties for researchers to understand differences between methods, it is less appropriate in relation to the end-user since the actual meaning of the class-labels is lost, and no clear statements about producer's and user's accuracy can be made.

(*b*) Fuzzy techniques for quality asessment. Gopal and Woodcock (1994) use fuzzy set theory in the accuracy assessment of thematic maps. The 'hard' results of the thematic map are analysed based on a 'soft' test data set and fuzzy operators. Expert knowledge is captured in a five point fuzzy membership scale, ranging from the linguistic *absolutely wrong* to *absolutely right*. The fuzzy operators allow to integrate information about the nature, frequency, magnitude, and source of errors in a thematic map.

The described fuzzy assessment technique could be of great benefit to assess the accuracy of a test data set based on expert information. However, it provides less information than the kappa statistics, since it only gives information on confusion between classes in the reference set, and not of those in the whole data set (map). Therefore the Kappa statistics are more appropriate for testing classifiers.

In a later paper, Woodcock (1996) points to the advantages that fuzzy set theory can have over the classical set theory in a classification process, which assumes each location in the map to be assigned unambiguously to a single class. He foresees possibilities for (future) improvement in the following three areas: (1) providing additional information on map accuracy; (2) expanding the kinds of queries that are possible with regard to spatial measurements on a map; (3) improved landscape characterization by explicit acknowledgment of the heterogeneity within the pixels.

(*c*) *Receiver operation characteristic* (*ROC*). In terms of certain features, the classification problem can be viewed as a detection problem. Kanungo *et al.* (1995) suggested an evaluation methodology that can be utilized to compare detection algorithms or to objectively evaluate the effect of parameter variation.

Briefly, the method involves calculation of the misdetection and false-alarm probabilities under different realizations and strengths of noise, added to two synthetic images differing only in the presence of a target. The different realizations of noise permit one to define for every signal to noise ratio (SNR) an operating characteristic, as the curve of P(misdetection)$=P$(no target/target) versus P(false alarm)$=P$(target/no target). Now, using the equal cost probability of error $P(E)$, one can plot $P(E)$ versus the SNR. A contrast threshold C_T is chosen at the value of SNR for a given $P(E)$, for instance, 0.25. In case the detection system is to be evaluated using different values of a variable of interest V, one should repeat the previous in order to obtain a curve of the variable of interest versus the contrast threshold.

Although this methodology, often referred to as the *receiver operating characteristic* (*ROC*), offers a way to compare target detection methods for robustness or to study their behavior under different parameter settings, it is fairly involved to apply this evaluation method to classification. The ROC has become a relatively common evaluation method in the medical field, for instance in computer-aided X-ray image analysis (Kanungo and Haralick 1995, Kobayashi *et al.* 1996).

(*d*) *Other techniques*. More *et al.* (1976) study the use of variance relationships among certain count estimators and introduce an estimator for the probability of correct classification, based on a stratified sampling scheme and posterior probabilities. Although the distinction they make between classified and unclassified observations is a rather academic one, computer simulations suggest the utility of the estimator. The authors do not report results on real data sets.

Rosenfield (1981) proposed the use of analysis of variance techniques. A major drawback for remote-sensing application is the assumption of independence between classes which in general does not hold for remote-sensing data.

Aronoff (1982b) presented a review paper on the theory and application of classification accuracy tests. He expresses the question of accuracy assessment in terms of hypothesis testing (H_0: the map is less accurate than required; H_1: the map accuracy is equal to or greater than required), and relates this to the terminology of the *producer's risk* and the *consumer's risk*. However, his method does not account for the large differences that may occur between the single class accuracies. In fact he concludes that confusion matrices may have a broader field of application.

Summarizing, various methods have been discussed that try to assess as objectively as is statistically possible the accuracy of classification results. The Kappa, KHAT, and Z statistics are considered the most appropriate techniques for reasons discussed before. The Kappa and KHAT statistics are suited for the description of a single classification result, and the Z statistic for the comparison between results.

Among the proposed approaches for accuracy assessment, these methods have the best mathematical foundation, and are thoroughly tested. In order to reduce as much as possible the bias introduced by the subjectivity in various stages of the assessment (e.g. training samples), we stress the importance of the analysis of more than one classification result by means of a pair-wise comparison using the multi-variate KHAT statistic (figure 5), that allows the converging line of evidence to emerge from the results.

4. The end-user's perspective

In the previous section, methodologies for assessing the *accuracy* and selecting the best among a given set of classifiers for a classification problem were outlined. Earlier it was observed that accuracy is hard to define, but the same holds for *quality*. Bratko *et al.* (1996), for instance, point out that in the evaluation of classifiers one has to consider the accuracy, the computational complexity, and the descriptional complexity of the generated results. In this section we will extrapolate the concept of accuracy to that of *quality*, being aware that we cannot do this except by creating the link between the accuracy and the user objectives and requirements (figure 6). Note that the term *user* in this context does not necessarily relate to the people who operate the image processing tools; the term refers to the actual user of the final product (e.g. the thematic map); hence the user objectives should allow an economical analysis of the value of the anticipated information.

This section addresses the issues of who the end-user really is (§4.1), what his requirements are (§4.2), and to what extent he is aware of the costs involved and how these costs can be fed back to the objectives (§4.3).

4.1. *Who is the end-user?*

Remote-sensing image interpretation is utilized in many applications, such as crop projections, planning, soil loss, water flow and use, urban development, and land-cover mapping. Therefore, rather than asking who the end-user is, it would be more appropriate to indicate on which application the attention is focused: for example, on land-cover mapping, and, more specifically, on agricultural applications.

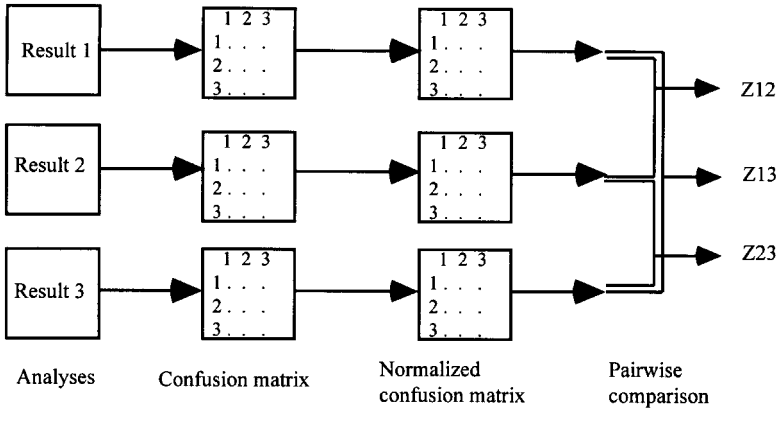

Figure 5. Reducing subjectivity in the assessment using various results and multi-variate statistical tools to select the best result (i.e. data set, feature vector, algorithm, or remote sensing system as a whole).

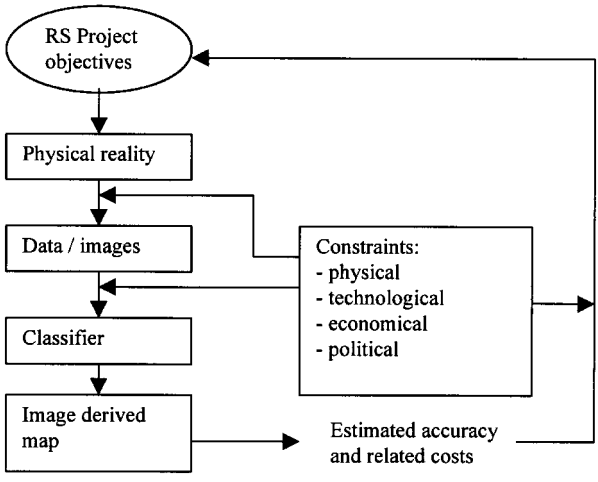

Figure 6. Interrelationship between the objective and different aspects of remote sensing derived maps.

Knowledge of land-cover is important for many planning and management activities concerned with the surface of the Earth. The term relates to the type of feature present on the surface of the Earth, such as trees, grasses, open water, and wetlands but does not assume the specific *use* of land. The term land use relates to the human activity associated with a specific piece of land, and may include characteristics that often cannot be inferred directly from the remote-sensing data alone.

Land-cover mapping has evolved in part with the data and analysis tools available. In the US, from the 1930s to the 1950s issues focused on the condition of the land resources (Lins and Kleckner 1996). In the 1960s and the early 1970s, the issues were those of land use monitoring and planning related to agriculture and urbanization. At that time, new source material from aerial photographs and satellite sensors was becoming available and a US national classification system was developed to exploit these sources. In the years following, the trend was toward environmental management which heavily depended on satellite remote-sensing and GIS. Agency or project-specific classification systems were designed to focus on ecologically based vegetation classification for land management. In the 1990s attention concentrated on monitoring the status and trends of the natural resources, both in the US and Europe. Assessments of ecosystems, biodiversity, and water quality required synoptic and repetitive coverage of the large areas to monitor changes in the landscape. This is illustrated by the various current US and European land-cover programs that extensively use land cover products like the US National Park Service Vegetation Mapping, US Coast Watch Change Analysis Project, and the European Community Framework Program Environment and Climate.

A special type of land-cover application can be found in agriculture. This application developed from aerial photo-interpretation for crop type classification, crop condition assessment, and crop yield estimation (Lillesand and Kiefer 1987) toward an advanced Agricultural Information System (AIS) integrating information of all kinds, ranging from crop calenders and soil temperature to imagery from sensors like the NOAA-AVHRR, ERS-1, and the Radarsat satellite. An important

characteristic of the AIS system is a highly automated land-cover map production. Such a system has been built in the framework of the Monitoring Agriculture with Remote-Sensing (MARS) Project (Vossen and Meyer-Roux 1995). The end-user in this case is the European Community DG-V (Agriculture).

4.2. *Does the user know what he wants or needs?*

Looking at land-cover applications in general, the user seems to know fairly well what he wants. According to the USGS, a land use/land cover classification system that can effectively employ orbital and high-altitude remote sensor data should meet the following criteria (Avery and Berlin 1985):

1. Interpretation accuracy in the identification of land use and land cover categories from remote sensor data should be 85% or greater.
2. The accuracy of interpretation for the several categories should be about equal.
3. Repeatable or repetitive results should be obtainable from one interpreter to another and from one sensing time to another.
4. The classification system should be applicable over extensive areas, which has implications for the speed of the system.
5. The categorization should permit vegetation and other types of land cover to be used as surrogates for activity.
6. The classification system should be suitable for use with remote sensor data obtained at different times of the year.
7. Effective use of sub-categories that can be obtained from ground surveys or from the use of larger-scale or enhanced remote sensor data should be possible.
8. Aggregation of categories must be possible.
9. Comparison with future land use data should be possible.
10. Multiple uses of land should be recognized when possible.

The above-listed requirements strongly relate to the key-issue addressed in this paper. The extent to which they are feasible is determined by the objectives set and the (physical) reality to which the requirements apply. Point 3, for instance, will be realistic only for a limited number of real-world applications, on the condition that the objectives, the specifications of the system, and the physical reality are tuned very well.

For land use/land cover data needed for planning and management, accuracy of interpretation at the generalized first and second levels (see also table 4) is satisfactory when the interpreter makes the correct interpretation 85 to 90% of the time (Avery and Berlin 1985). Recently, the USGS and the US Environmental Protection Agency (USEPA) executed an extended survey of requirements for land cover/use data (Lins 1994). One of the objectives of the survey was to gather information from as many data users as possible and then to identify ways to reduce expenditures for collecting data, minimize duplication, maximize benefits, identify standards for data accuracy and transferability, and make data more readily available.

In the survey, 35% of the respondents regarded a classification accuracy of more than 90% critical for their application (55% of the respondents regarded it desirable, 2% not necessary). If an opportunity for data sharing and exchange were to be created, about one third of the respondents regarded a classification accuracy of 95% necessary. Some relevant results are reported (percentages may total more than

Table 4. Representative image interpretation formats for various land use/land cover classification levels (modified after Lillesand and Kiefer 1987, Avery and Berlin 1985). The use of satellite imagery in level III is possible for a limited number of classes with specific knowledge based image classification.

Classification level	Representative map compilation scale	Sensor platform or altitudes	Representative format for image interpretation
I	1:500 000	Earth satellites	Landsat-MSS, ERS-1
II	1:62 000	9000–22 000 m Earth satellites	Small scale aerial photography, Landsat-TM, SPOT, and IRS-1 images, ERS-1, Radarsat
III	1:24 000	3000–9000 m Earth satellites	Medium scale aerial photography, EarlyBird, QuickBird, Eyeglass
IV	1:10 000	1200–3000 m	Large scale aerial photography

100% because of multiple responses) in table 5. Regarding the requirements for level of detail, the response of table 6 was reported. For agriculture, no explicit statements were found about the accuracy requirements related to classification. However, conventional methods to control farmer's declarations have an accuracy of 1/10 ha.

For an AIS to be useful, it appears that the applied image classification algorithms should have real-time capabilities and should be able to distinguish between about 20 different crop-types. It is noted that previously this would require medium scale aerial photography (table 4). However, recently, in many cases, the data from modern space-borne sensors combined with specific agricultural knowledge can suffice.

4.3. *…and what that costs?*

Of course, the user of classification results has an awareness of costs. In fact, a user decides to use remote-sensing tools to reduce the costs and time of sending people in the field to obtain the desired information. Trade-offs to reduce cost are the very *raison d'etre* of the remote-sensing community. Moreover, no one has unlimited budgets, which a-priori imposes restrictions on the type and number of

Table 5. User indicated importance of classification accuracy (after Lins 1994).

Classification accuracy (%)	Frequency	Percentage (%)
95	123	31
90	226	57
85	40	10
80	13	3
75	3	1

Table 6. User indicated importance of level of detail (after Lins 1994).

Level of detail	Frequency	Percentage (%)
Minimal processing	17	4
Level I	21	5
Level II	77	19
Level III	208	52
Greater detail	91	23

data sets used. The issue of cost reduction has been one of the items addressed in the survey of Lins (1994). Table 7 shows the areas in which one would make a compromise if trade-offs had to be made to reduce cost.

However, it remains interesting to see how far this awareness goes. Are the needs formulated in the previous section feasible? What is the price for the requirements? To this end, we distinguish between two types of costs that are important for a remote-sensing based land-cover mapping project: those related to the remote-sensing data, and those induced by the error of classification into the final product. The cost effectiveness of the application of remote-sensing in monitoring agriculture, for instance, is very sensitive to the costs of the imagery, which at present are very high.

5. Proposed cost-based evaluation

In order to obtain an idea of the cost related to the classification result, the following scenario is proposed, based on the findings in the previous sections:

1. Define precise objectives and determine a cost table, which relates to each misclassification at certain costs.
2. Collect a number of classification results with sufficiently accurate ground truth information (see §3.1).
3. Determine for each analysis the confusion matrix, normalize it, and estimate Kappa and the variance with equation (1) and (2).
4. Use the Z statistics (3) to determine which approach is significantly better.
5. Compute the application cost for the best classifier, and test whether these costs meet the criteria set in the objectives.

This cost-based approach applied to a hypothetical classifier performance evaluation in the next subsection.

5.1. *An illustration: cost of error in agricultural monitoring due to classification errors*

The consequences of error in classification results can be illustrated by means of a simple numerical example of control with remote-sensing of arable and land subsidies, in which remote-sensing can serve as a verification tool for farmer area-payment declarations. In 1996 the US New Farm Legislation became effective, which practically eliminated the deficience payment rates as known before (USDA 1996). However, in Europe crop subsidies are still effective, which justifies the example. In such a concrete example, the errors of omission and commission become less abstract.

Table 7. Areas of compromise to reduce costs.

Item	Frequency	Percentage (%)
Decrease frequency of update	213	54
Cost share	210	53
Lower spatial resolution	119	30
Increase min map unit	119	30
Reduce class detail	89	22
Limit coverage area	89	22
Map to source material	55	14
Dec resolution of source	47	12
Lower class accuracy	49	12
No trade-offs	22	6

Table 8. Hypothetical subsidies for four crops.

Category index/name	Subsidy [\$ ha^{-1}]
A Cereals	$s_A = 10$
B Oil seeds	$s_B = 14$
C Protein crops (1)	$s_C = 12$
D Protein crops (2)	$s_D = 10$

For instance, errors of omission can result in categories not being detected and subsidies will be reduced: the farmers pay for this type of error. The errors of commission, on the other hand, can overestimate certain categories and may therefore lead to over-subsidy, paid by the state or the tax payer.

In this example, the five steps described in the previous subsection will be followed:

1. *Define precise objectives and determine the cost matrix* (table 8). Assuming that in a certain area of interest, three crops are of particular interest to a subsidy provider: cereals, oil seeds, and protein crops. The subsidies related to these crops may vary like those in table 8. The user now defines or estimates the costs for mixing when farmer area payment declarations have to be checked. For instance, one could think of the following simple confusion cost matrix, in which the costs of misclassification are directly related to differences in subsidies:

$$\begin{bmatrix} 0 & |s_A - s_B| & |s_A - s_C| & |s_A - s_D| \\ |s_B - s_A| & 0 & |s_B - s_C| & |s_B - s_D| \\ |s_C - s_A| & |s_C - s_B| & 0 & |s_C - s_D| \\ |s_D - s_A| & |s_D - s_B| & |s_D - s_B| & 0 \end{bmatrix} = \begin{bmatrix} 0 & 4 & 2 & 0 \\ 4 & 0 & 2 & 4 \\ 2 & 2 & 0 & 2 \\ 0 & 4 & 2 & 0 \end{bmatrix} \quad (7)$$

   Note that in this example it is just accidental that confusion-cost matrix is symmetric.
2. *Collect a number of classification results.* For illustration, in figure 7 (*b*, *c*, and *d*) three different classification results are reported using as input data bands *B7*, *B9* and *B10* of a Daedalus 1268 Airborne Thematic Mapper (ATM) scanner.
3. *Determine the confusion matrices* (tables 9–11) and from them the normalized confusion matrices (tables 12–14).
4. *Using Z statistics, determine which approach is significantly better.*

K1 = 0.8689	sigma1 = 2.2978	Z12 = − 0.0122
K2 = 0.9325	sigma2 = 4.6874	Z13 = 0.1039
K3 = 0.6204	sigma3 = 0.6641	Z23 = 0.0659

5. *Compute the cost of applying the best classifier.* The process of computing the cost of a classification result needs as input: the normalized confusion matrix, the confusion-cost matrix, and the classification result. In the following it is shown how this information is combined to generate an estimate of the *cost of error*. First, a weighted cost matrix is produced by an element-by-element

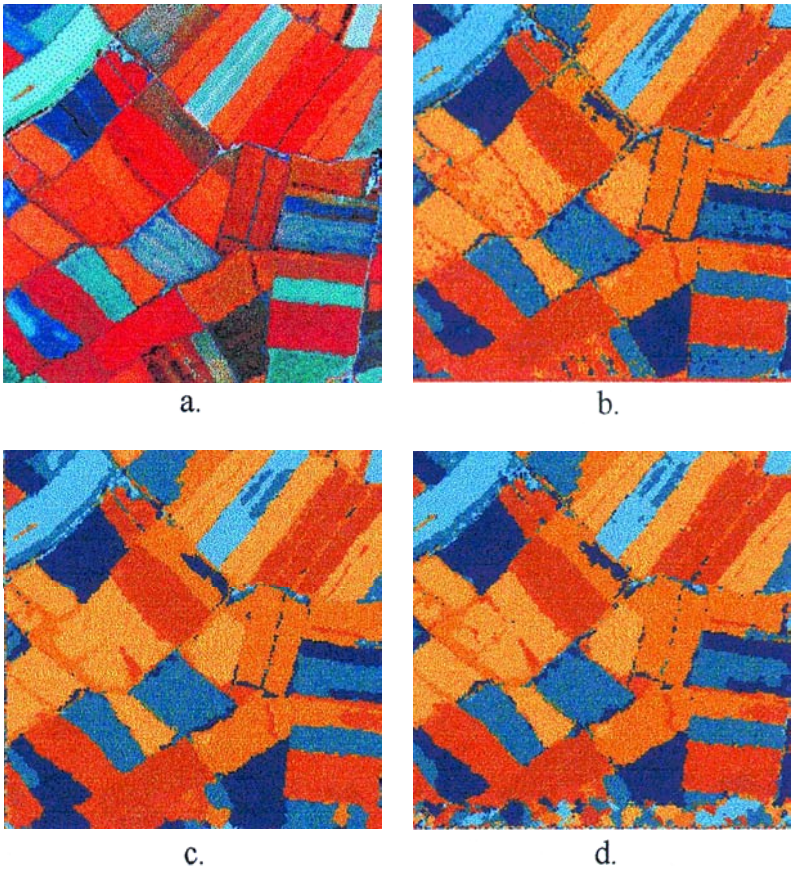

Figure 7. Examples of three different classification results. (*a*) False colour composite image of the ATM bands 7 (red), 9 (blue) and 10 (green) of a Feltwell region (UK). (*b*) Maximum Likelihood classification (classifier 1). (*c*) and (*d*) Two different implementations of a Markov Random Field method (classifier 2 and 3, respectively).

multiplication of the normalized confusion matrix (considered as a measure for classifier performance) and the confusion cost matrix:

$$\underbrace{\begin{bmatrix} p(y=A/x=A) & \dots & p(y=D/x=A) \\ \vdots & \ddots & \vdots \\ p(y=A/x=D) & \dots & p(y=D/x=A) \end{bmatrix}}_{\text{Classifier performance}} \text{Element by element} \underbrace{\begin{bmatrix} 0 & \dots & |s_A - s_D| \\ \vdots & 0 & \vdots \\ |s_D - s_A| & \dots & 0 \end{bmatrix}}_{\text{Confusion cost table}}$$

$$= \underbrace{\begin{bmatrix} c(y=A\,|x=A) & c(y=B\,|x=A) & c(y=C\,|x=A) & c(y=D\,|x=A) \\ c(y=A\,|x=B) & c(y=B\,|x=B) & c(y=C\,|x=B) & c(y=D\,|x=B) \\ c(y=A\,|x=C) & c(y=B\,|x=C) & c(y=C\,|x=C) & c(y=D\,|x=C) \\ c(y=A\,|x=D) & c(y=B\,|x=D) & c(y=C\,|x=D) & c(y=D\,|x=D) \end{bmatrix}}_{\text{Weighted cost}} \tag{8}$$

Using the outcomes of step 1 and 3 yields the following weighted cost table:

0	0.2991	0.0164	0
0.1712	0	0.1605	0.1847
0	0.1890	0	0
0	0	0.0126	0

In order to obtain the *absolute cost* matrix, this weighted cost matrix is now matrix-multiplied by the matrix of area, in which every row r is filled with the estimate of area A_r for class r. The area estimates of the different classes can either be derived from the classification result by summing up the pixels belonging to the

Table 9. Confusion matrix for classifier 1.

	A	B	C	D	Total
A	35	5	1	0	41
B	1	34	6	1	42
C	0	2	35	0	37
D	2	0	1	4	7
	38	41	43	5	127

Table 10. Confusion matrix for classifier 2.

	A	B	C	D	Total
A	37	5	0	0	42
B	0	33	2	1	36
C	0	2	35	1	38
D	0	0	1	47	48
	37	40	38	49	164

Table 11. Confusion matrix for classifier 3.

	A	B	C	D	Total
A	0	2	4	2	8
B	1	33	4	5	43
C	0	10	25	7	42
D	3	8	4	29	44
	4	53	37	43	137

Table 12. Normalized confusion matrix for classifier 1.

	A	B	C	D
A	0.9170	0.0748	0.0082	0
B	0.0428	0.8307	0.0802	0.0462
C	0	0.0945	0.9053	
D	0.0402	0	0.0063	0.9538

Table 13. Normalized confusion matrix for classifier 2.

	A	B	C	D
A	1.0000	0.0157	0	0
B	0	0.9221	0.0471	0.0271
C	0	0.0622	0.9175	0.0151
D	0	0	0.0354	0.9578

Table 14. Normalized confusion matrix for classifier 3.

	A	B	C	D
A	0.8508	0.0264	0.0861	0.0368
B	0.0442	0.6783	0.1341	0.1434
C	0	0.1651	0.6736	0.1613
D	0.1050	0.1302	0.1062	0.6585

respective classes, or by multiplying the prior probabilities of the classes by the total observed area.

$$\underbrace{\begin{bmatrix} c(y=A/x=A) & \dots & c(y=D/x=A) \\ \vdots & \ddots & \vdots \\ c(y=A/x=D) & \dots & c(y=D/x=D) \end{bmatrix}}_{\text{Weighted cost}} \underbrace{\begin{bmatrix} A_A & A_A & A_A & A_A \\ A_B & A_B & A_B & A_B \\ A_C & A_C & A_C & A_C \\ A_D & A_D & A_D & A_D \end{bmatrix}}_{\text{Estimated area per category}}$$

$$= \underbrace{\begin{bmatrix} C_{A/A} & \dots & C_{D/A} \\ \vdots & \ddots & \vdots \\ C_{A/C} & \dots & C_{D/D} \end{bmatrix}}_{\text{Estimated 'absolute' cost}} \tag{9}$$

When analysing a remote-sensing image of 40 km × 40 km or 160.000 ha, and supposing that the respective prior probabilities are, $p(x=A)=0.05$, $p(x=B)=0.25$, $p(x=C)=0.30$, and $p(x=D)=0.05$, the estimated area matrix becomes

8000	8000	8000	8000
40 000	40 000	40 000	40 000
48 000	48 000	48 000	48 000
8000	8000	8000	8000

and the absolute confusion cost matrix becomes (prices in \$10.000)

1.2751	1.2751	1.2751	1.2751
1.0550	1.0550	1.0550	1.0550
0.7561	0.7561	0.7561	0.7561
0.0603	0.0603	0.0603	0.0603

Summing up the costs over all the elements in the matrix yields the total amount of cost of error, which \$125.860, amounting to about 10% of the budget for subsidizing $0.65 \times 160\,000$ ha (with an average of \$12 ha^{-1} = \$1 248 000). In this case it may be wise to decide to use a better classifier, a more precise data set, or to use another method.

Assuming a somewhat better classifier (or data set, or remote-sensing system as a whole) that has a performance matrix as shown below, the total cost of error would easily be reduced to less than 5% of the budget.

Many comments can be made regarding the given example. For instance, many assumptions may be too simplistic, or exaggerated. However, it is thought that the concept of relating the outcome of the classifiers to the objectives defined by the end-users has been made clear: a precise idea about the costs involved in classification error, whatever the reason, may cast a different light on the usefulness of remote-sensing image classification products for a specific application. It will help to identify the most cost-effective type of data, or of the classification procedure, and helps end-users to justify their claim for accuracy requirements.

6. Conclusion

In this paper the quality assessment land-cover maps derived from remote-sensing images has been discussed. Three issues relate to the subject of quality assessment: the classification methods as such, the methods to evaluate the classification results, and the requirements.

Some problems that come with image classification algorithms in general have been discussed. A number of evaluation methods has been reviewed, and it is concluded that the KHAT analysis is the most suited one if one is interested in comparing classifiers. This comparison of classifiers is regarded fundamental in order to eliminate to a certain extent the subjectivity that is inevitably introduced in the evaluation process.

Regarding the requirements, it has been shown in a survey (Lins 1994) that for many remote-sensing projects a class-accuracy of 90% is critical. However, this number can only be justified if end-users connect this requirement to their objectives. This paper proposes a protocol for assessing the quality of classification results, related to the economical aspects of a remote-sensing project.

It was shown, using an example of monitoring agriculture with remote-sensing, that in certain cases the economic cost of misclassification can be high, and that it may be advantageous for the user to reconsider either the objectives, the type of data used, or other aspects of the remote-sensing system that he uses to produce the map.

The proposed protocol gives a clear and to the point answer whether a classification result meets the objectives, and is proposed as an important step toward a standard in the remote-sensing information processing flow.

Acknowledgments

This work was supported by the European Community Program Training and Mobility for Researchers (Marie Curie Fellowship) under Contract ERBFMBICT950257. It was performed in part during two visits by the first author to the Digital Image Analysis Lab, University of Arizona, Tucson (AZ). The authors wish to thank the following persons for their kind support and valuable comments: Dr Steve Yool, Department of Geography and Regional Development, University

of Arizona, Tucson (AZ), Jennifer Dungan JCWS, Inc./NASA Ames Research Center, Moffett Field (CA), Susan Benjamin, USGS/NASA Ames Research Center, Moffett Field (CA), and Dr Russel Tronstad, Department of Agricultural/Resource Economics, University of Arizona, Tucson (AZ).

Thanks are also due to Hunting Technical Services Ltd. for providing the ATM images of the Feltwell region.

References

ARONOFF, S., 1982a, Classification accuracy: a user approach. *Photogrammetric Engineering and Remote Sensing*, **49,** 1299–1307.

ARONOFF, S., 1982b, The map accuracy report: a user's view. *Photogrammetric Engineering and Remote Sensing*, **48,** 1309–1312.

AVERY, T. E., and BERLIN, G. L., 1985, *Fundamentals of remote-sensing and airphoto interpretation* (New York: Macmillan Publishing Company).

BRATKO, I., CESTNIK, B., and KONONENKO, I., 1996, Attribute-based learning. *AI Communications*, **9,** 27–32.

CARD, D. H., 1982, Using known map category marginal frequencies to improve estimates of thematic map accuracy. *Photogrammetric Engineering and Remote Sensing*, **48,** 431–439.

COHEN, J., 1960, A coefficient of agreement for nominal scales. *Educational Psychology Measurement* , **20,** 37–46.

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

CONGALTON, R. G., and MEAD, R. A., 1983, A quantitative method to test for consistency and correctness in photointerpretation. *Photogrammetric Engineering and Remote Sensing*, **49,** 69–74.

CONGALTON, R. G., ODERWALD, R., and MEAD, R., 1983, Landsat classification accuracy using discrete multivariate analysis statistical techniques. *Photogrammetric Engineering and Remote Sensing*, **49,** 1671–1678.

CZAPLEWSKI, R. L., 1992, Accuracy assessment of remotely sensed classification with multiphase sampling and the multivariate composite estimator. *Proceedings of the 16th International Biometrics Conference*, 7–11 December, 1992, pp. 2–22.

DUIN, R. P. W., 1996, A note on comparing classifiers. *Pattern Recognition Letters*, **17,** 529–536.

ESTES, J. E., and MOONEYHAN, D. W., 1994, Of maps and myths. *Photogrammetric Engineering and Remote Sensing*, **60,** 517–524.

FIENBERG, S. E., 1970, An iterative procedure for estimation in contingency tables. *The Annals of Mathematical Statistics*, **41,** 907–917.

FIENBERG, S. E., and HOLLAND, P. W., 1970, Methods for eliminating zero counts in the contingency tables. In *Random Counts in Scientific Work*, edited by G. P. Patil (University Park, Pennsylvania: Pennsylvania State University Press), pp. 233–260.

FITZPATRICK-LINS, K., 1981, Comparison of sampling procedures and data analysis for a land-use and land-cover map. *Photogrammetric Engineering and Remote Sensing*, **47,** 343–351.

FOODY, G. M., and COX, D. P., 1994, Subpixel land cover composition estimation using a linear mixture model and fuzzy membership functions. *International Journal of Remote Sensing*, **15,** 619–631.

FU, K. S., and MUI, J. K., 1981, A survey on image segmentation. *Pattern Recognition*, **13,** 3–16.

GINEVAN, M. E., 1979, Testing land-use map accuracy: another look. *Photogrammetric Engineering and Remote Sensing*, **45,** 1371–1377.

GOPAL, S., and WOODCOCK, C., 1994, Theory and methods for accuracy assessment of thematic maps using fuzzy sets. *Photogrammetric Engineering and Remote Sensing*, **60,** 181–188.

HARALICK, R. M., and SHAPIRO, L. G., 1993, *Computer and Robot Vision* (New York: Addison-Wesley).

HARALICK, R. M., and SHAPIRO, L. G., 1985, Survey: image segmentation techniques. *CVGIP*, **29,** 100–132.

HARDIN, P. J., 1994, Parametric and nearest-neighbor methods for hybrid classification: a comparison of pixel assignment accuracy. *Photogrammetric Engineering and Remote Sensing*, **60,** 1439–1448.

HAY, A. M., 1979, Sampling designs to test land-use map accuracy. *Photogrammetric Engineering and Remote Sensing*, **45,** 529–533.

HORD, R. M., and BROONER, W., 1976, Land-use map accuracy criteria. *Photogrammetric Engineering and Remote Sensing*, **42,** 671–677.

HUDSON, W. D., and RAMM, C. W., 1987, Correct formulation of the Kappa Coefficient of agreement. *Photogrammetric Engineering and Remote Sensing*, **53,** 421–422.

JAIN, R. C., and BINFORD, T. O., 1991, Ignorance, myopia and naiveté in computer vision systems. *CVGIP Image Understanding*, **53,** 112–117.

JANSSEN, L. L. F., and VAN DER WEL, F. J. M., 1994, Accuracy assessment of satellite derived land-cover data: a review. *Photogrammetric Engineering and Remote Sensing*, **60,** 419–426.

JENSEN, J. R., 1996, *Introductory Digital Image Processing—A Remote Sensing Perspective* (Englewood Cliffs, New Jersey: Prentice-Hall).

KANUNGO, T., JAISIMHA, M. Y., PALMER, J., and HARALICK, R. M., 1995, A methodology for quantitative performance evaluation of detection algorithms. *IEEE Transactions on Image Processing*, **4,** 1667–1673.

KANUNGO, T., and HARALICK, R. M., 1995, Receiver operating characteristic curves and optimal Bayesian operating points. *Proceedings of International Conference on Image Processing, Washington, 23–26 Oct. 1995*, pp. 256–259.

KEW, N. R., 1996, Information-theoretic measures for assessment and analysis in image classification. In *Soft computing in remote-sensing data analysis*, edited by E. Binaghi, P. A. Brivio and A. Rampini (Singapore: World Scientific), pp. 173–180.

KOBAYASHI, T., XIN-WEI, X. U., MACMAHON, H., METZ, C. E., and DOI, K., 1996, Effect of a computer-aided diagnosis scheme on radiologists' performance in detection of lung nodules on radiographs. *Radiology*, **199,** 843–848.

KONONENKO, I., and BRATKO, I., 1991, Information based evaluation criterion for classifier's performance. *Machine Learning*, **6,** 67–80.

LANDGREBE, D. A., and MALARET, E. R., 1986, Noise in remote sensing systems: effect on classification accuracy. *IEEE Transactions on Geoscience and Remote Sensing*, **GE-24,** 294–299.

LATHAM, J. P., 1963, Methodology for an instrumented geographic analysis. *Ann. Assoc. Amer. Geogr.*, **53,** 194–209.

LEE, C., and CHEN, L., 1995, High-speed closest codeword search algorithms for vector quantization. *IEEE Transactions on Communications*, **43,** 323–331.

LILLESAND, T. M., and KIEFER, R. W., 1987, *Remote Sensing and Image Interpretation* (New York: John Wiley & Sons).

LINS, K., 1978, Accuracy and consistency comparisons of land use and land cover maps made from high-altitude photographs and Landsat multispectral image. *Journal of Research of the US Geological Survey*, February.

LINS, K., 1994, Requirements analysis results for land cover and land use data. *US Geological Survey*, March.

LINS, K. F., and KLECKNER, R. L., 1996, Land use and land cover mapping in the United States: an overview and history of the concept. In *Gap Analysis, A Landscape Approach to Biodiversity Planning*, edited by J. M. Scott, T. H. Tear, and F. W. Davis (Bethesda: American Society for Photogrammetry and Remote Sensing), pp. 57–67.

LUNETTA, R. S., CONGALTON, R. G., FENSTERMAKER, L. K., JENSES, J. R., MCGWIRE, K. C., and TINNEY, L. R., 1991, Remote-sensing and geographic information system data integration: error sources and research issues. *Photogrammetric Engineering and Remote Sensing*, **57,** 667–687.

MA, Z., and REDMOND, R. L., 1995a, Analysis of classification results of remotely sensed data and evaluation of classification algorithms. *Photogrammetric Engineering and Remote Sensing*, **61,** pp. 427–433.

MA, Z., and REDMOND, R. L., 1995b, Tau coefficients for accuracy assessment of classification of remote-sensing data. *Photogrammetric Engineering and Remote Sensing*, **61,** 435–439.

MALING, D. H., 1989, *Measurements from maps: principles and methods of cartometry* (Oxford: Pergamon Press), pp. 144–177.

MATHER, P. M., 1987, Computer processing of remotely-sensed images: an introduction (New York: John Wiley & Sons).

MAXIM, L. D., and HARRINGTON, L., 1983, The application of pseudo-bayesian estimators to remote-sensing data: ideas and examples. *Photogrammetric Engineering and Remote Sensing*, **49,** 649–658.

MORE, D. S., WHITSITT, S. J., and LANDGREBE, D. A., 1976, Variance comparisons for unbiased estimators of probabilities of correct classifications. *IEEE Transaction on Information Theory*, **IT-22,** 102–105.

RICHARDS, J. A., 1993, *Remote-sensing digital image analysis* (New York: Springer-Verlag).

RICHARDS, J. A., 1996, Classifier performance and map accuracy. *Remote Sensing of Environment*, **57,** 161–166.

ROSENFIELD, G. H., 1981, Analysis of variance of thematic mapping experiment data. *Photogrammetric Engineering and Remote Sensing*, **47,** 1685–1692.

ROSENFIELD, G. H., FITZPATRICK-LINS, K., and LING, H. S., 1982, Sampling for the thematic map accuracy testing. *Photogrammetric Engineering and Remote Sensing*, **48,** 131–137.

ROSENFIELD, G. H., and MELLEY, M., 1980, Applications of statistics to thematic mapping. *Photogrammetric Engineering and Remote Sensing*, **46,** 1287–1294.

SHAHSHAHANI, B. M., and LANDGREBE, D. A., 1994, The effect of unlabeled samples in reducing the small sample size problem and mitigating the Hughes phenomenon. *IEEE Transactions on Geoscience and Remote Sensing*, **32,** 1087–1095.

STEHMAN, S. V., 1996, Estimating the Kappa coefficient and its variance under stratified random sampling. *Photogrammetric Engineering and Remote Sensing*, **62,** 401–407.

STORY, M., and CONGALTON, R. G., 1986, Accuracy assessment: a user's perspective. *Photogrammetric Engineering and Remote Sensing*, **52,** 397–399.

STRAHLER, A. H., 1981, Stratification of natural vegetation for forest and rangeland inventory using Landsat digital imagery and collateral data. *International Journal of Remote Sensing*, **2,** 15–41.

SWAIN, P. H., 1978, Fundamentals of pattern recognition in remote-sensing. In *Remote-sensing: The Quantitative Approach*, edited by P. H. Swain and S. M. Davis (New York: McGraw-Hill), pp. 136–187.

THOMAS, I. L., and ALLCOCK, G. MCK., 1984, Determining the confidence level for a classification. *Photogrammetric Engineering and Remote Sensing*, **50,** 1491–1496.

UNITED STATES DEPARTMENT OF AGRICULTURE, 1996, USDA 1996 Farm Bill Fact Sheet, April 1996.

VAN GENDEREN, J. L., LOCK, B. F., and VASS, P. A., 1978, Remote sensing: statistical testing of thematic map accuracy. *Remote Sensing of Environment*, **7,** 3–14.

VOSSEN, P., and MEYER-ROUX, J., 1995, Crop monitoring and yield forecasting activities of the MARS Project. In *European Land Information Systems for Agro-Environmental Monitoring*, edited by D. King, R. J. A. Jones, and A. J. Thomasson, EUR Publication N. 16232/EN: 11–30. More references and information regarding the MARS project can be found via the Space Application Institute webpage: http://www.jrc.it/.

WOODCOCK, C. E., 1996, On roles and goals for map accuracy assessment: a remote-sensing perspective. *Proceedings of the Second International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences*, edited by H. T. Mowrer, R. L. Czaplewski, and R. H. Hamre (Fort Collins, Colorado: USDA Forest Service Rocky Mountain Research Station), pp. 535–540.

ZAMPERONI, P., 1996, Plus ça va, moins ça va. *Pattern Recognition Letters*, **17,** 671–677.

ZHU, Z., OHLEN, D. O., CZAPLEWSKI, R. L., and BURGAN, R. E., 1996, Alternative method to validate the seasonal land cover regions of the conterminous United States. In *Proceedings of the Second International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences*, edited by H. T. Mowrer, R. L. Czaplewski and R. H. Hamre (Fort Collins, Colorado: USDA Forest Service Rocky Mountain Research Station), pp. 409–418.

ZHUANG, Y. J., 1996, A survey on evaluation methods for image segmentation. *Pattern Recognition* , **29,** 1335–1346.

ZHUANG, X., ENGEL, B. A., XIONG, X., and JOHANNSEN, C. J., 1995, Analysis of classification results of remotely sensed data and evaluation of classification algorithms. *Photogrammetric Engineering and Remote Sensing*, **61,** 427–433.