Accuracy Assessment of Satellite Derived Land-Cover Data: A Review

Lucas L. F. Janssen and Frans J. M. van der Wel

Abstract

Accuracy assessment of land-cover classifications derived from remote sensing data has been recognized as a valuable tool in judging the fitness of these data for a particular application. Recent research initiatives in the area of spatial data accuracy and integration of remote sensing data in geographic information systems have revived the discussion on accuracy assessment. This article aims at contributing to this discussion by means of a review based on a division into positional and thematic accuracy.

An important observation is that there are a limited number of methods for assessing data accuracy. However, the applied definitions differ very much from author to author, especially in the assessment of thematic accuracy.

Accuracy assessment mostly yields one single measure such as root-mean-square error or proportion of pixels correctly classified. These single measures do not give sufficient information and they can be based on statistically or methodologically non-valid methods. Therefore, not a single measure but also the total process of assessing these measures should explicitly be reported.

Introduction

The accuracy of spatial data has been defined by the United States Geological Survey as: "The closeness of results of observations, computations, or estimates to the true values or the values accepted as being true" (USGS, 1990).

Accuracy assessment or validation is an important step in the processing of remote sensing data. It determines the value of the resulting data to a particular user, i.e., the information value. During the last two decades, a large number of articles has been published on accuracy assessment of land-cover classifications. Very different approaches for validation have been presented, usually with a particular application in mind for the data in hand.

At present, the geographic information systems (GIS) and remote sensing communities pay more and more attention to accuracy topics. Technological developments in the area of data processing offer more and more possibilities. A responsible use of the stored geodata is only possible if the quality of these data is known. Furthermore, integrated processing of different types of geodata can be performed much more responsibly if (again) the quality of the data is known.

Studies on spatial data accuracy and on errors in the integration of remote sensing data in a GIS have been initiated by the National Center for Geographic Information Analysis (NCGIA). Their Research Initiatives 1 (Accuracy of Spatial Databases) and 12 (Integration of Remote Sensing and GIS Technologies) will contribute to the understanding of the magnitude and attitude of errors in remote sensing data and how they affect accuracy (Goodchild and Gopal, 1989; Lunetta *et al.*, 1991).

In this article, a review of existing accuracy assessment procedures for land cover classifications, derived from data acquired by land observation satellites such as the Landsat and SPOT series, will be given. Methods applied for both positional and thematic accuracy are mentioned and commented on. At this point it is important to stress that there are differences in meaning among the following three concepts: precision, accuracy, and reliability. They all have a clear meaning which is explicitly stated for both positional and thematic characteristics. In this way, an attempt has been made to reduce the confusion which undoubtedly exists in this field and to contribute to the ongoing discussion.

Although the word "map" can still be found in many articles and presentations, in this article the word map should be interpreted as "digital raster data" as derived from image processing. The cartographic presentation of digital data is a scientific field in itself and is not dealt with in this article.

Remote Sensing and Land Cover

Until now, land-cover and land-use data have been mainly acquired from terrestrial surveying and visual aerial photointerpretation. Photointerpretation is based on human vision and pattern recognition capacities. Identification of terrain objects is based on nine interpretation keys: pattern, tone, texture, shadow, site, shape, size, association, and resolution (for the interested reader, a standard text on these topics such as Avery and Berlin (1985) is recommended).

The interpretation process can be facilitated by viewing the photographs stereoscopically. Air photointerpretation keys also assist the interpreter by offering guidelines for the identification of certain information classes (Lillesand and Kiefer, 1987; p. 115). Objects are distinguished by a combination of both geometric and thematic properties. A good example is the delineation of individual trees in a forest stand.

As a result of an interpretation process, a "representation of the world" is obtained, consisting of terrain objects with a geometric and a thematic component. Therefore, both

L.L.F. Janssen is with DLO—Winand Staring Centre for Integrated Land, Soil and Water Research, P.O. Box 125, 6700 AC Wageningen, The Netherlands.

F.J.M. van der Wel is with the University of Utrecht, Faculty of Geographical Sciences—Cartography Section, P.O. Box 80.115, 3508 TC Utrecht, The Netherlands.

Photogrammetric Engineering & Remote Sensing, Vol. 60, No. 4, April 1994, pp. 419–426.

0099-1112/94/6004–419$03.00/0

visual photo-interpretation and terrestrial surveying are typically directed to vector based data of terrain objects describing land cover or land use.

Remote sensing is a data acquisition technique. Earth observation satellites such as Landsat and SPOT measure the relative amount of electromagnetic radiation that is reflected (and emitted) by the Earth's surface. In fact, this is a sampling process dividing the Earth's surface into equal areas called scene elements. The corresponding image representation of a scene element is known as a picture element or pixel. The measurements of these elements in several spectral bands are converted and stored in a limited number of quantization levels (e.g., 8- or 16-bits code). The stored values are referred to as digital numbers (DN).

A remote sensing image can be characterized by an image space and a feature space. The position of a pixel represented in the image space is determined by a unique row and column index (i,j). The relative spectral reflection values $(DN_1, \ldots, DN_n)$ can be represented in the n-dimensional feature space.

In most projects remote sensing images undergo two transformations:

- a registration of the image coordinate system into a certain map projection, enabling other geodata to be used; and
- a classification of the continuum of spectral data into nominal user-desired classes (the most subjective transformation).

The classification or interpretation of remote sensing images can be performed in a visual and a digital way. Visual interpretation offers more or less the same characteristics and properties as visual photo interpretation. Until now, most digital interpretation has been based solely on the per-pixel multivariate data. These per-pixel classifications are limited to the interpretation element "tone" as used in visual interpretation. This limitation has two major implications:

- per-pixel classifications by definition yield spectral classes mainly related to land cover, where land use is mainly determined from contextual and associative information. Campbell (1987; p. 473) puts it as follows: "land cover designates the visible evidence of land use, to include both vegetative and nonvegetative features."
- per-pixel classifications yield thematic information per raster element. When looking at a classification result, although one can distinguish fields, for instance, it should be noted that terrain objects as such are not explicitly stored. The raster data derived from remote sensing should be considered as point data that have a certain spatial extent.

Positional Accuracy

Definitions

By and large, positional accuracy of remote sensing data refers to the accuracy of a geometrically rectified image. Rectification includes registration to a reference coordinate system together with a resampling procedure where in this context the term georeferencing and geocoding are used (Irish, 1990). Georeferencing means that a link between an image and a reference coordinate system is established (registration). Geocoding implies that an image is also resampled into a new raster format. We will use registration to indicate the geometrical link between geodata stored in different coordinate systems.

Accuracy assessment of the registration of images can benefit from experiences already available in photogrammetry. This discipline has led to the development of specific measurement methods and has a characteristic vocabulary.

Precision relates to the exactness with which a certain coordinate can be determined, for example, the explicitness of "pointing precision." Precision depends on the method applied and the characteristics of the data. It is quantified by giving an indication of the performance of a particular measurement when several repetitions are required and is therefore usually defined in terms of standard error. Precision does not reveal the absolute closeness to the "correct" coordinate. This absolute closeness is indicated by *accuracy*. For instance, if Ground Control Points (GCPs) are used for registration, the accuracy is partly determined by the pointing precision of the GCPs in both the image and reference coordinate system. *Reliability* describes the possibilities of statistical detection of gross and possibly systematic errors occuring during a geometric correction (Molenaar, 1980). Internal and external reliability refer to input and output data, respectively.

Registration

In a registration, the row-column image data are related to the coordinate system of another image data or to a particular map projection system. This relationship can be determined in two ways:

- from the orbital parameters of the satellite, or
- by locating identical ground control points (GCPs) in both the image and reference coordinate system.

The latter non-parametric approach is generally accepted as the most realistic option because the orbital geometry model used to describe the errors is incomplete and causes geometric distortions (Mather, 1987; p. 130). The registration accuracy can be derived from the registration process itself because the selection of GCPs is performed in a relatively objective way.

GCPs are points that can be well identified in both the image coordinate system (source coordinates i,j) and in a reference coordinate system (reference coordinates x,y). These coordinate systems can be related by using polynomial equations. The complexity or order of these polynomials depends on the geometry of the image and the type of map projection. A first-degree affine transformation is often sufficient for satellite images.

The polynomial is calculated by means of a minimization of "the sum of squares." After determination of the optimal solution, the residuals in both the x and y directions (δ_x, δ_y) can be calculated. Then these residuals are used for accuracy assessment by calculating a root-mean-square (RMS) error or standard deviation. The RMS error in the x- direction is calculated as

$$RMS_x = \left[\frac{1}{n}\sum_{i=1}^{n}(\delta_{x_i})^2\right]^{1/2}$$

where δ_{x_i} = the residual of the i^{th} GCP and n = the number of GCPs.

The RMS error in the y direction is calculated similarly and subsequently the RMS_x and RMS_y can be combined to yield one planimetric RMS error (RMS_{xy}):

$$RMS_{xy} = [RMS_x^2 + RMS_y^2]^{1/2}$$

A statistically more sound estimation of the accuracy would be to calculate a standard deviation. In this case the sum of the residuals is divided by the redundancy (r) which

depends on the degree of freedom determined by the applied polynomial

$$s_x = \left[\frac{1}{r}\sum_{i=1}^{n}(\delta_{xi})^2\right]^{1/2}$$

If a large number of GCPs are used, then the RMS error and standard deviation will converge.

The meaning of the calculated accuracies can be understood by constructing confidence limits. If the GCPs are independently identified, one can construct confidence limits by using a Gaussian approach, thereby assuming that the residuals are distributed normally. An example is shown in Figure 1. The consequenses of positional uncertainty should not be forgotten when overlay operations are performed, as with cross tabulation or multi-temporal classification.

Accurate identification of GCPs is a prerequisite for obtaining an accurate registration. The derivation of these points from maps can introduce an amount of uncertainty because a map represents an idealization and generalization of "reality." To eliminate this uncertainty, Global Positioning System (GPS) techniques could be used. Nevertheless, the surplus value of these techniques for a more precise identification is bound by the GCPs determined from the remote sensing image. The identification possibilities are largely determined by the spatial resolution of the applied scanner.

The residuals calculated in the registration can be used for statistical testing. Buiten (1988) presented a variance-ratio and data-snooping test. In the variance-ratio test the estimated variances of the total number of GCPs in the *x*- and *y*-direction are tested against an a priori variance. This *a priori* variance is determined by the identification, digitization, and source accuracy of the GCPs. The variance-ratio test evaluates the hypothesis that the transformation model is relevant and that no gross errors are introduced during the selection of GCPs. If the hypothesis is rejected, the data-snooping test evaluates the error for every single GCP.

Resampling

The registration process has been explained in the previous section. After registration, the image can be combined with other geo-information. Most often, the image is also resampled into a new grid that corresponds to the *x*- and *y*-axis of the chosen reference coordinate system.

Resampling can be performed for original spectral data as well as for classified (nominal) data. There are three resampling methods currently being used: nearest neighbor, bilinear interpolation, and cubic convolution. The applied resampling method should correspond to the character of the data (continue or nominal) and the relationship of the original to the resulting pixel size. Subsampling (resampled pixel size smaller than original size) can give a future user wrong ideas about the spatial resolution.

Resampling of spectral data, however, also means a transformation of the feature space. Assumptions concerning normal distribution of spectral classes should be checked when spectral data are subsampled and subsequently classified. Smith and Kovalick (1985) compared the effects that different resampling techniques have on classification results. They concluded that it does not make much difference if nearest neighbor is performed before or after classification.

Observations and recommendations

The registration of satellite images is relatively straightforward. Positional accuracy defines the relationship between the registered image and the applied source data (map). The spatial resolution of the present land observation satellites circumscribes the identification precision of GCPs and therefore the potential registration accuracy.

Preferably, a registered image should be labeled with information on the following items: source of reference GCPs, number of GCPs, type of transformation, RMS error or standard deviation, and, if necessary, resampling method.

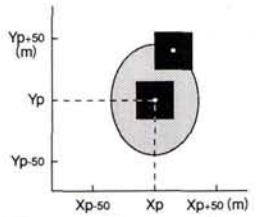

Figure 1. The coordinates for a geocoded pixel are (Xp,Yp); the pixel size is 30 by 30 m². Registration yielded standard deviations of s_x = 18 m and s_y = 23 m. The hatched part indicates the positions to which the pixel center (Xp,Yp) refers (confidence level 95 percent).

Thematic Accuracy

Definitions

Thematic accuracy refers to the non-positional characteristics of a spatial data entity, the so-called attributes (Chrisman, 1987). If the attribute permits some classification, this kind of accuracy is often termed *classification accuracy* (Hord and Brooner, 1976). In remote sensing classifications, this accuracy refers to the correspondence between the class label assigned to a pixel and the "true" class. The true class can be observed in the field directly or indirectly; for example, from a reference map.

Aronoff (1989) is right to emphasize the statistical meaning of accuracy. Consequently, he defines classification accuracy as: "... the probability that the class assigned to a location on the map is the class that would be found at that location in the field ... ." Story and Congalton (1986) refer to this accuracy as "user's accuracy" or *reliability*, being a measure of the value of a map for a particular user.

Attribute *precision* is a somewhat less known quality of spatial data and, therefore, it deserves some attention. Although Aronoff (1989) considers precision to be a component of positional accuracy, attribute precision refers to the repeatability of class assignments, i.e., the agreement of a series of repeated identifications of the same entity. In this sense, automatic clustering (applying one parameter-set) would yield a very precise classification result. In fact, "uncertainty" might be a better name for the quality that Aronoff distinguishes; if detail increases, the possibility of errors occurring increases too, meaning more uncertain data.

Classified Data	Reference Data X	Y	Z	Row total	Correct (%)
X	24	2	4	30	80
Y	6	45	9	60	75
Z	3	5	52	60	87
Column total	33	52	65	150	

Figure 2. The error matrix in a lay-out as presented by Story and Congalton (1986). Numbers in the matrix represent numbers of pixels. For each class the correct percentage is added.

Campbell (1987) defines attribute precision as "detail," or the number of classes identified during a classification procedure. It refers to the generalization level of the classification. According to this definition, precision refers to only one observation. However, the generally accepted meaning of precision is related to the performance of a method (repeatability), indicating the involvement of more observations.

Methods Applied

The assessment of positional accuracy is based on analysis of the residuals calculated in a registration. Likewise, the thematic accuracy of a classification could be based on the Euclidian or statistical distances calculated in the classification itself. However, the training stage in classification is very subjective compared to the identification of ground control points. Therefore, the distances calculated in a classification cannot be considered independent and useful for accuracy assessment.

Thematic accuracy is most often assessed by evaluating a sample population of the classification result. The sample should be taken randomly. In practice, however, due to time and money constraints, random sampling proves to be a problem. The classes that are determined by classification of a remote sensing image are compared to reference data originating from field survey, aerial photographs, or other digital geodata. Comparison is made by establishing an "error matrix," which yields information on the accuracy of individual classes as well as the accuracy of the classification as a whole. Congalton (1991) gives a review of methods for comparison of error matrices.

Sampling and Field Survey

The sample to be drawn consists of a number of sampling units. These sampling units can be points, lines, or areas. The choice of the sampling unit is very important. For example, areas are often used to evaluate land-cover "map" accuracy (Dicks and Lo, 1990). We would like to stress that remote sensing data should be considered to be "point-sampled" data, in which the points possess a certain spatial extent. From a theoretical point of view, individual pixels are the most appropriate sampling units if a per-pixel classification is performed. In some cases, e.g., when applying spatial smoothing, a cluster-based sampling is most appropriate (Todd and Gehring, 1980). Poor accessibility of the terrain and limited budgets may also result in the application of cluster-based sampling.

Preferably, stratified sampling should be used and should be based on the distinguished classes. In practice, there is often a time lapse between the acquisition date and the classification date. In these cases, only a spatial random distribution can be used for sampling as it cannot be based on the distribution of the individual classes.

One has to decide whether or not to incorporate positional uncertainty in the field survey and subsequent cross-tabulation of the reference and classified data. If one wants to make a clear distinction between positional and thematic accuracy, positional uncertainty should be taken into account. This can be achieved by involving the contextual information in the identification of the true land-cover type of a pixel.

Without departing from the subject, it must be stated that "truth" has a certain subjective dimension. A remote sensing image is initially classified into spectral classes (land cover). Of course, one can compare this result with geodata based on land use (functionality) but at the same time it can be expected that the remote sensing classification will yield bad results.

Congalton (1988) stresses the importance of the decisions made in the sampling procedure. After all, it is assumed that these determine the value of the derived accuracy as representative estimates of the accuracy of the entire area under consideration. For a thorough discussion on sampling techniques, readers are referred to Cochran (1977).

Error Matrix

The comparison or cross-tabulation of the classified land cover to the actual land cover revealed by the sample sites results in an *error matrix, confusion matrix, contingency table* (Story and Congalton, 1986), *evaluation matrix* (Aronoff, 1984), or *misclassification matrix* (Chrisman, 1991). We will use "error matrix" to indicate the summarized sample results. Different measures and statistics can be derived from the values in an error matrix. In this article, several approaches will be examined. The non-statistical measures are described in this section, while procedures based on the binomial distribution and those based on coefficients of agreement are described in the previous sections.

Although the matrix is simple in itself, there is some confusion regarding the "lay-out." This might seem rather trivial, but the consequences can be considerable: different meanings of rows and columns in matrices undoubtedly obstruct a straightforward interpretation. *Classified* (Story and Congalton, 1986; Mather, 1987), *predicted* (Hay, 1979), *evaluated* (Campbell, 1987), *interpreted* (Van Genderen *et al.*, 1978), or *observed* (Aronoff, 1982b) data all indicate the same concept. Similarly, do *reference* (Congalton *et al.*, 1983; Story and Congalton, 1986), *verified* (Aronoff, 1982b), *identified* (Hay, 1979), *known* (Lillesand and Kiefer, 1987), or *true* (Card, 1982) data or class. Curran (1985) and Aronoff (1982b) even use the identical term (*observed*) to indicate reference and classified data, respectively! The lay-out presented by Story and Congalton (1986) in their description of the error matrix will be adopted here (Figure 2).

Once the error matrix is established, a number of accuracy measures can be derived. Again, authors use various terms for the same error type. Most important is the meaning of the calculated measure. The Proportion of pixels Correctly Classified (PCC) (Veregin, 1989) is calculated by dividing the number of correctly classified samples positioned at the diagonal of the matrix (Figure 2) by the total number of pixels checked. This value is a measure of the classification as a whole. The PCC calculated from Figure 2 is (24 + 45 + 52) divided by 150 equals 81 percent.

Different measures of individual classes can be calculated using two approaches:

- user's and producer's accuracy (Story and Congalton, 1986), and
- errors of omission and commission.

The user's accuracy, calculated as the number of correctly classified samples divided by the row total, provides the user information about the accuracy of the land-cover data. A user's accuracy of 80 percent for class X means that 80 percent of the pixels classified as X are X in reality. Then, the accuracy based on the sampled pixels is representative of the total classification result. User's accuracy is sometimes called "reliability."

Dividing the number of correctly classified samples by the column total yields the producer's accuracy: it indicates the percentage of samples of a certain (reference) class that were correctly classified. There are many examples of remote sensing studies in which these accuracies have been calculated (e.g., Prisley and Smith, 1987; Felix and Binney, 1989). Because both the user's and producer's accuracy can be of interest for a certain user of the data, these terms are deceiving. Therefore, it seems favorable to express thematic accuracy in terms of "error of omission" and "error of commission."

In Figure 3, it is explained how errors of omission and commission are calculated. Errors of omission refer to the samples of a certain class of the reference data that were not classified as such. Errors of commission refer to the samples of a certain class of the classified data that were wrongly classified. Both types of errors are very important in the iterative training stage because they indicate which categories have unrepresentative distributions. Campbell (1987) gives a rather extended explanation of the error matrix, but unfortunately he does not set a good example; he calculates the errors of commission by using the total number of samples from the reference instead of those from the classification!

As can be deduced from the above descriptions, the following relationships are valid:

- user's accuracy (%) = 100 % − error of commission (%)
- producer's accuracy (%) = 100 % − error of omission (%)

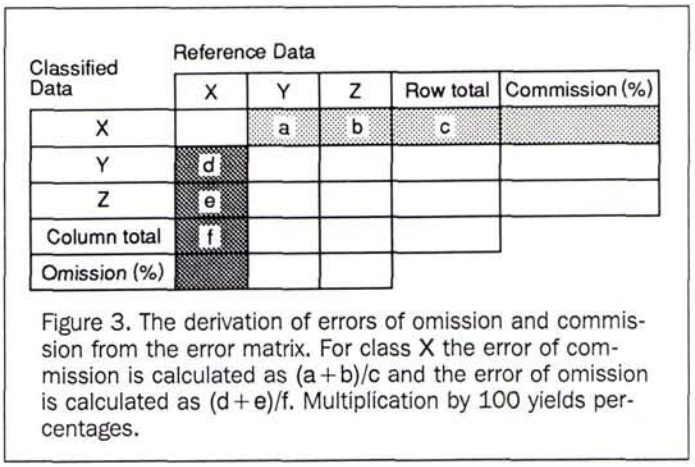

Figure 3. The derivation of errors of omission and commission from the error matrix. For class X the error of commission is calculated as (a+b)/c and the error of omission is calculated as (d+e)/f. Multiplication by 100 yields percentages.

Procedures Based on the Binomial Distribution

In the previous section, the figures calculated are often taken as representative of the total classification result. If appropriate sampling is performed, confidence limits can be determined and hypothesis testing can be carried out. The statistics can be calculated for both the individual categories and the classification as a whole. The examples in this section are based on the error matrix presented (Figure 2).

Some accuracy assessment methods use the binomial distribution as an approximation to the hypergeometric distribution, which is exact for finite populations (the number of pixels is always finite). The binomial model distinguishes between correct and incorrect samples (e.g., Davis, 1986). Binomial probabilities can be calculated either from the binomial probability density function itself, or derived from the normal approximation (Aronoff, 1982a).

Confidence Limits

Confidence intervals can be calculated for the PCC from the sample size n, the number of correct classifications k, and significance level α. The 95 percent confidence interval for the PCC can be read from binomial nomograms as given in statistical handbooks, or calculated using the exact binomial distribution. The upper and lower limit for the PCC are 73.4 percent and 86.7 percent, respectively, for $n = 150$, $k = 121$, and $\alpha = 0.05$.

In addition to the discrete binomial distribution, the continuous normal approximation to the binomial has also been used to calculate confidence limits of the estimated - population - accuracy. Rosenfield and Melley (1980) describe the correction required for this adjustment. This approximation is valid if the sample sizes are large; this prerequisite is not always satisfied, as Ginevan (1979) noted, referring to Hord and Brooner (1976). The latter describe a procedure by which the 95 percent confidence interval can be derived, given the sample size and PCC, and suggest that only the lower limit be used. Using Hord and Brooner's approach, one would find a lower and upper limit for the PCC of 73.9 percent and 86.5 percent, respectively, for $n = 150$, $k = 121$, and $\alpha = 0.05$.

The PCC confidence limits assessment shows that the calculated PCC is the center of an interval in which the actual PCC can be found with $1-\alpha$ confidence.

Hypothesis Testing

For some applications, the classification result should have a minimum PCC. In these cases, hypothesis testing is most appropriate. Hypothesis testing of a predetermined accuracy is generally applied in quality control. Acceptance sampling is a topic of quality control. The advantage of the acceptance sampling approach is that the minimum sample size can be determined if risks and predetermined accuracy are defined. For a comprehensive treatise on topics on statistical quality control in general and acceptance sampling in particular, see Grant and Leavenworth (1988).

In hypothesis testing, the following parameters have to be defined:

- null (H_0) and alternative (H_1) hypothesis, and
- significance level α.

The significance level defines the possibility of wrongly rejecting H_0 (type I error). Optionally, the power of the test ($1 - \beta$) can be defined as the possibility of wrongly accepting H_0 (type II error) for situations that are valid under H_1.

It should be realized that there is a certain asymmetry in testing. H_0 is accepted unless it is significantly proven that it should be rejected. This mechanism is shown in the next

two tests, both of which could be used to test a minimum required accuracy of 80 percent:

Test Definition 1:

H_0: classification accuracy $\geq$ = 80 %

H_1: classification accuracy < 80 %

under the conditions of α = 0.05 and n = 150. The binomial distribution can be used to calculate the critical values: the outcomes of k that reject H_0. For this situation the critical values are in the interval [0,111].

Test Definition 2:

H_0: classification accuracy < 80%

H_1: classification accuracy $\geq$ 80%

under the conditions of α = 0.05 and n = 150. For this situation the critical values are in the interval [129,150].

If a person that ordered a remote sensing classification wishes to get a result with a minimum accuracy, the second test definition is the most effective. Then, the burden of evidence is on the producer of the data to prove that the accuracy is at least 80 percent.

In general, the terms consumer's and producer's risk are used to indicate α and β, respectively. Obviously, the consumer's risk defines the risk of wrongly accepting H_0, which has the largest consequences for the consumer. Note that the terms producer's and consumer's *risk* have a completely different meaning than do producer's and user's *accuracy*.

Ginevan (1979) used the binomial probability density function in an acceptance sampling procedure to calculate the optimal sample size before actual sampling, thereby minimizing the field survey. Ginevan's approach is the following: if the required accuracy and α and β are defined, the optimal sample size n in combination with the maximum allowable misclassifications can be calculated. Given a required accuracy of 85 percent, a consumer's risk of 0.05, and a producer's risk of 0.043 (for a classification with an actual accuracy of 95 percent), the optimal sample size is 93 with a maximum allowed number of misclassifications of 8. Using the same table, the optimal sample size of a classification with an actual accuracy of 90 percent (same consumer's risk), the optimal sample size would be much larger than 400. Minimizing the producer's risk can be clearly balanced against the number of samples and the related field survey. Because Ginevan (1979) did not present a table for a required accuracy of 80 percent, it is impossible for us to elaborate this approach on our example.

Aronoff (1982a) considered the procedure worked out by Ginevan (1979) as statistically valid and emphasized the importance of including both consumer's and producer's risks in an accuracy assessment procedure. In fact, both risks should be minimal, which is difficult because of their interdependency; a smaller producer's risk can be obtained by increasing the consumer's risk or increasing the sample size.

Aronoff (1985) introduced the minimum accuracy value. During an accuracy test, instead of rejecting a classification as insufficiently accurate, the highest accuracy for which the number of misclassifications would indeed pass the test can be assessed. Aronoff's approach is the following: given a required accuracy of 80 percent and a consumer's risk of 0.05, we would find that our example (n = 150 and 29 misclassifications, Figure 2) does not pass the test. At the same time, it can be found that the minimum accuracy is 74.0 percent and that the producer's risk is 0.04 for a classification with an actual accuracy of 90 percent. The minimum accuracy value is a useful index for comparing accuracy tests with different sample sizes.

Coefficients of Agreement

The measures described in previous sections reflect the closeness of the result compared to the truth. Another objective of accuracy assessment can be to compare different classification results to test the effectiveness of a certain classifier or the ancillary data applied. PCC values cannot be compared in a straightforward way and therefore other methods are described. A solution would be to normalize the error matrices (Congalton *et al.*, 1983).

Another approach is to calculate the Kappa-coefficient, which may be used to compare different classification methods that are based on the same data (Congalton and Mead, 1983; Congalton *et al.*, 1983). Congalton *et al.* (1983) introduced the Kappa-coefficient of agreement as an accuracy measure for remote sensing classifications. Kappa takes the chance agreement into account; as stated by Campbell (1987), Kappa: "... adjusts the percentage correct measure by subtracting the estimated contribution of chance agreement ... ."

The Kappa-coefficient for the error matrix in Figure 2 results in a value of 0.70. This implies that the accuracy of the classification is 70 percent better than the accuracy that would result from a random assignment. Calculation of Kappa involves the complete error matrix, including information concerning errors of omission and commission. The exact formulation is decribed by Hudson and Ramm (1987). Rosenfield and Fitzpatrick-Lins (1986) suggest using Kappa as a sort of standard measure of accuracy for thematic classifications as a whole and propose a coefficient of conditional Kappa for individual classes.

Observations and Recommendations

Especially in the assessment of thematic accuracy, we found the terminology for error matrices, the derived types of error, and the risks in hypothesis testing to be abundant. We think that some standardization would be favorable.

If the error matrix is based on simple random sampling of individual pixels, there are appropriate techniques, using the binomial distribution, for determination of confidence limits, hypothesis testing, or determination of the optimal sample size. However, in a large number of remote sensing studies, we discovered that no appropriate random sampling had been applied or that the per-pixel classification results were compared to a database consisting of polygons. In the latter case, point observations were compared to an interpretation result that had been derived by generalization and idealization of the truth. It is not always fully understood that this approach conceptually differs very much from point sampling. Therefore, more attention should be given to the characteristics of the method of sampling and its effect on estimated accuracy.

Due to a number of reasons, accuracy assessment of thematic data differs from project to project. Therefore, it is important to give ample information on the complete validation procedure. An answer should be given to the following questions:

- which sampling strategy and which type of sampling units were chosen ?
- is the positional uncertainty of remote sensing and reference data taken into account?

  and
- which error measures were calculated, which assumption

were made, and to what extent can they be held representative for the classification result as a whole?

Concluding Remarks

A large number of methods and definitions are used to descibe the accuracy of land-cover data derived from remote sensing data. Moreover, different methods have been developed for positional and thematic accuracy assesment because registration and classification are completely independent.

The registration accuracy of satellite images can be based on the residuals that were calculated during the registration. Because of the relatively objective identification of GCPs, the residuals are suitable for this purpose. Although registration can be more difficult for relieved terrains, the selection of GCPs is rather objective and measures can be calculated from the registration process itself.

The assessment of thematic accuracy is much more complex. The training stage in a classification is very subjective. Therefore, the distance measures as calculated during the classification cannot be used for accuracy assessment. Because of this, thematic accuracy should be assessed by comparing a sample of the classification result with reference data. In this article we suggested that results of a per-pixel classification should be considered as point classifications, and that validation should preferably be based on the sampling of individual pixels. Because of the abundance of terms in remote sensing literature on the subject of thematic accuracy, the meaning of the terms applied should be explicit.

References

Aronoff, S., 1982a. Classification Accuracy: A User Approach. *Photogrammetric Engineering & Remote Sensing* 48(8):1299–1307.

———, 1982b. The Map Accurcay Report: A User's View. *Photogrammetric Engineering & Remote Sensing* 48(8):1309–1312.

———, 1984. An Approach to Optimized Labeling of Image Classes. *Photogrammetric Engineering & Remote Sensing* 50(6):719–727.

———, 1985. The Minimum Accuracy Value as an Index of Classification Accuracy. *Photogrammetric Engineering & Remote Sensing* 51(1):99–111.

———, 1989. *Geographic Information Systems: A Management Perspective*. WDL Publications, Ottawa. 294 p.

Avery, T. E., and G. L. Berlin, 1985. *Interpretation of Aerial Photographs*. Fourth Edition. Burgess, Minneapolis. 554 p.

Buiten, H. J., 1988. Matching and mapping of remote sensing images: aspects of methodology and quality. *Proceedings 16th ISPRS-Congress*, Kyoto, Japan, July 1988, 27-B10(III):321–330.

Campbell, J. B., 1987. *Introduction to Remote Sensing*. The Guilford Press, New York, London. 551 p.

Card, D. H., 1982. Using Known Map Category Marginal Frequencies to Improve Estimates of Thematic Map Accuracy. *Photogrammetric Engineering & Remote Sensing* 48(3):431–439.

Chrisman, N. R., 1987. The Accuracy of Map Overlays: A Reassessment. *Landscape and Urban Planning* 14:427–439.

———, N. R., 1991. The Error Component in Spatial Data. *Geographical Information Systems* (D. J. Maguire, M. F. Goodchild, and D. W. Rhind, editors), Longman Scientific & Technical, pp. 165–174.

Cochran, W. G., 1977. *Sampling Techniques*. Wiley & Sons, New York. 428 p.

Congalton, R. G., 1988. A Comparison of Sampling Schemes Used in Generating Error Matrices for Assessing the Accuracy of Maps Generated from Remotely Sensed Data. *Photogrammetric Engineering & Remote Sensing* 54(5):593–600.

———, 1991. A Review of Assessing the Accuracy of Classifications of Remotely Sensed Data. *Remote Sensing of Environment* 37:35–46.

Congalton, R. G., and R. A. Mead, 1983. A Quantitative Method to Test for Consistency and Correctness in Photointerpretation. *Photogrammetric Engineering & Remote Sensing* 49(1):69–74.

Congalton, R. G., R. G. Oderwald, and R. A. Mead, 1983. Assessing Landsat Classification Accuracy Using Discrete Multivariate Analysis Statistical Techniques. *Photogrammetric Engineering & Remote Sensing* 49(12):1671–1678.

Curran, P. J., 1985. *Principles of Remote Sensing*. Longman, London, New York. 282 p.

Davis, J. C., 1986. *Statistics and Data Analysis in Geology*. Second Edition. Wiley & Sons, New York. 646 p.

Dicks, S. E., and T. H. C. LO, 1990. Evaluation of Thematic Map Accuracy in a Land-Use and Land-Cover Mapping Program. *Photogrammetric Engineering & Remote Sensing* 56(9):1247–1252.

Felix, N. A., and D. L. Binney, 1989. Accuracy Assessment of a Landsat-Assisted Vegetation Map of the Coastal Plain of the Arctic National Wildlife Refuge. *Photogrammetric Engineering & Remote Sensing* 55(4):475–478.

Ginevan, M. E., 1979. Testing Land-Use Map Accuracy: Another Look. *Photogrammetric Engineering & Remote Sensing* 45(10):1371–1377.

Goodchild, M. F., and S. Gopal, 1989. *Accuracy of Spatial Databases*. Taylor & Francis, London, New York, Philadelphia. 290 p.

Grant, E. L., and R. S. Leavenworth, 1988. *Statistical Quality Control*. McGraw-Hill International Editions, New York. 714 p.

Hay, A. H., 1979. Sampling Designs to Test Land-Use Map Accuracy. *Photogrammetric Engineering & Remote Sensing* 45(4):529–533.

Hord, R. M., and W. Brooner, 1976. Land-Use Map Accuracy Criteria. *Photogrammetric Engineering & Remote Sensing* 42(5):671–677.

Hudson, W. D., and C. W. Ramm, 1987. Correct Formulation of the Kappa Coefficient of Agreement. *Photogrammetric Engineering & Remote Sensing* 53(4):421–422.

Irish, R., 1990. Geocoding Satellite Imagery for GIS Use. GIS World, August/September, pp. 59–62.

Lillesand, T. M., and R. W. Kiefer, 1987. *Remote Sensing and Image Interpretation*. Wiley & Sons, New York. 721 p.

Lunetta, R. S, R. G. Congalton, L. K. Fenstermaker, J. R. Jensen, K. C. McGwire, and L. R. Tinney, 1991. Remote Sensing and Geographic Information System Data Integration: Error Sources and Research Issues. *Photogrammetric Engineering & Remote Sensing* 57(6):677–687.

Mather, P. M., 1987. *Computer Processing of Remotely-Sensed Images. An Introduction*. Wiley & Sons, Chichester. 352 p.

Molenaar, M., 1985. Quality Evaluation of Photogrammetric Point Determination. *Photogrammetria* 40:165–177.

Prisley, S. P., and J. L. Smith, 1987. Using Classification Error Matrices to Improve the Accuracy of Weighted Land-Cover Models. *Photogrammetric Engineering & Remote Sensing* 53(9):1259–1263.

Rosenfield, G. H., and K. Fitzpatrick-Lins, 1986. A Coefficient of Agreement as a Measure of Thematic Classification Accuracy. *Photogrammetric Engineering & Remote Sensing* 52(2):223–227.

Rosenfield, G. H., and M. L. Melley, 1980. Applications of Statistics to Thematic Mapping. *Photogrammetric Engineering & Remote Sensing* 46(10):1287–1294.

Smith, J. L., and B. Kovalick, 1985. A Comparison of the Effects of Resampling before and after Classification on the Accuracy of a Landsat Derived Cover Type Map. *Proc. Intern. Conf. on the RS Society and the Center for Earth Resources Management*, University London, September, pp. 391–400.

Story, M., and R. G. Congalton, 1986. Accuracy Assessment: A

User's Perspective. *Photogrammetric Engineering & Remote Sensing* 52(3):397–399.

Todd, W. J., and D. G. Gehring, 1980. Landsat Wildland Mapping Accuracy. *Photogrammetric Engineering & Remote Sensing*, 46(4):509–520.

U.S. Geological Survey, 1990. *The Spatial Data Transfer Standard*, Draft, January 1990.

Van Genederen, J. L., B. F. Lock, and P. A. Vass, 1978. Remote Sensing: Statistical Testing of Thematic Map Accuracy. *Remote Sensing of Environment* 7:3–14.

Veregin, H., 1989. *A Taxonomy of Error in Spatial Databases*. Technical Paper 89-12, NCGIA, Santa Barbara. 113 p.

(Received 10 March 1992; revised and accepted 21 January 1993; revised 16 February 1993)