Proceedings of the 8th International Symposium on
Spatial Accuracy Assessment in Natural Resources and Environmental Sciences
Shanghai, P. R. China, June 25-27, 2008, pp. 154-162

Sample Size Determination for Image Classification Accuracy Assessment and Comparison

Giles M. Foody [+]

School of Geography, University of Nottingham, NG7 2RD, UK

Abstract. The classification accuracy statement is the basis of the evaluation of a classification's fitness for purpose. Accuracy statements are also used for applications such as the evaluation of classifiers, with attention focused especially on differences in the accuracy with which data are classified. Many factors influence the value of a classification accuracy assessment and evaluation programme. This paper focuses on the size of the testing set(s), and its impacts on accuracy assessment and comparison. Testing set size is important as an inappropriately large or small sample could lead to limited and sometimes erroneous assessments of accuracy and of differences in accuracy. In this paper the basic statistical principles of sample size determination are outlined. Some of the basic issues of sample size determination for accuracy assessment and accuracy comparison are discussed. With the latter, the researcher should specify the effect size (minimum meaningful difference), significance level and power used in an analysis and ideally also fit confidence limits to estimates. This will help design a study as well as aid interpretation. In particular, it will help avoid problems such as under-powered analyses and provide a richer information base for classification evaluation. Central to the argument is a discussion of Type II errors and their control. The paper includes equations that could be used to determine sample sizes for common applications in remote sensing, using both independent and related samples.

Keywords: remote sensing classification, sample size, Type I and II error, power, confidence interval.

1. Introduction

The results of an accuracy assessment may be used in a variety of ways. Fundamentally, accuracy assessment is typically undertaken in remote sensing studies to provide an objective basis on which to evaluate the quality of the thematic map represented by the classification output. However, aspects of accuracy assessment may also underpin and inform other activities. Frequently, the accuracy statement derived for a classification is used in evaluating the classifier used, with the relative superiority of classifiers evaluated on the accuracy with which they classify data [1].

This article focuses on a key issue in the design of an accuracy assessment programme, testing set size. The latter is important as the use of an excessively large sample may lead the conclusion that any non-zero difference observed is statistically significant [2]. Conversely, the use of a sample size that is too small may result in a failure to detect a difference that may actually be large and important. This article draws on the published literature to highlight especially concerns over the power of analyses used to detect differences and flag the value of power analyses and confidence limits in enhancing the interpretation of results.

2. Basic issues in sample size determination for accuracy assessment

The most widely used methods of accuracy assessment are site-specific approaches based on the entries in a confusion matrix. One of the most popular measures of classification accuracy is the proportion of correctly allocated cases. Since it is typically impractical to evaluate the accuracy of the entire area mapped, the assessment of classification accuracy is typically based on a sample of cases (e.g. pixels) drawn from it. This sample is the testing set. To make credible generalizations to the entire map from this sample, it is important

[+] *E-mail address*: giles.foody@nottingham.ac.uk

ISBN: 1-84626-171-6, **ISBN13:** 978-1-84626-171-8

that the testing set be acquired following an appropriate sampling design. The latter typically involves some form of random sampling in order to provide an unbiased and representative sample that is suitable for the purpose of generalization. One key issue in the design of an accuracy assessment programme is the size of the sample that forms the testing set.

It is important to design the accuracy assessment programme carefully to meet its aims. Care should be taken to ensure that the sample size used is appropriate, noting that both an unduly large or small sample size may have negative impacts on a study. A variety of approaches have been used to determine the required sample size [3]. One popular approach is to use basic sampling theory to derive an estimate of the required sample size. This may be approached from a precision- or a power-based perspective. With the former, the aim is essentially to estimate accuracy, expressed as the proportion of correctly allocated cases, to a certain degree of precision. With this approach the researcher is essentially setting the width of the confidence interval around the estimated accuracy. The latter may be expressed as

$$p \pm h = p \pm z_{\alpha/2}(\mathrm{SE}) \tag{1}$$

where p is the observed proportion of correctly allocated cases which is an estimate of the actual population value P, h the half width of the desired confidence interval and $z_{\alpha/2}$ the critical value of the normal distribution for the two-tailed significance level α and SE is the standard error of the estimate. The latter is

$$\mathrm{SE} = \sqrt{\frac{p(1-p)}{n}} \tag{2}$$

where n is the sample size. These equations may be rewritten to provide an estimate of the sample size required to estimate the population proportion. Indeed, the required sample size to estimate the population (entire mapped area) proportion of correctly allocated cases in a classification may be estimated from

$$n = \frac{z^2_{\alpha/2} P(1-P)}{h^2} \tag{3}$$

where P is a planning value for the population proportion. This approach has been used to define the size of the testing set in remote sensing studies. The approach has the advantage of being based on established statistical theory and provides the researcher with a means to design the sampling to satisfy fundamental requirements. For simplicity, it will be assumed throughout this paper that simple random sampling is used to acquire the testing set and that accuracy is to be expressed by measures based on the proportion of correctly allocated cases. Although many other approaches are adopted, this is a popular basis for accuracy assessment and many other approaches (e.g. use of stratified random sample designs) are well-established adaptations of this approach.

Although the use of equation 3 for sample size estimation is based on established and accepted theory it may not, however, always be appropriate. The use of the equation requires specification three terms, the half width of the confidence interval, a planning value for P and a significance or confidence level that determines the magnitude of $z_{\alpha/2}$. It may not always be easy to select appropriate values for these terms. Some values may be selected following conventional practice. For example, a 0.05 significance level is typically adopted by convention and so $z_{\alpha/2}$ =1.96. Some of the values may be estimated from prior knowledge or conservative values selected. For example, in the absence of knowledge on a planning value for P, an estimate, if potentially conservative one (large), may be derived using P=0.5 as at this value the term $P(1\text{-}P)$ is maximised. However, it is often unclear what value to use for h.

A further problem with using equation 3 for sample size determination is that it simply gives the sample size required to estimate a proportion at a given level of precision which is only part of the common accuracy assessment scenario encountered in many remote sensing studies. Implicitly at least, there is often a desire to evaluate the derived proportion against some other value, often a target value, to determine if the classification satisfies the user's specified accuracy needs. For example, a target accuracy of 0.85 (or 85%) is often suggested in remote sensing applications, although the use of this value is debatable [4]. If such a target value was selected, the analyst might use the testing set, defined with the aid of equation 3, to estimate the proportion of correctly allocated cases and compare its lower confidence limit, specified at an appropriate significance level, to the target value. If the lower confidence limit exceeds the target value then the classification could be taken to meet the user's needs. Clearly, if the classification accuracy was only just

above the target value a very narrow confidence interval would be required if the classification was to be deemed as acceptable and this would necessitate the use of a very large testing sample. It would, however, be very easy to use a sample size that was too large or too small. The use of the basic confidence interval based approaches for sample size determination will often result in the use of a sample that is too small to detect a meaningful difference; an under-powered analysis. It may be preferable to design the sample more fully around the factors that influence the comparison of proportions as this allows the required sample size to be determined to meet project goals and for which there is an extensive literature for guidance. The aim, therefore, is not simply to estimate the proportion of correctly allocated cases but to compare it against another proportion, usually a target value or one derived from the application of another classification analysis. In such circumstances, further issues require consideration and these, unfortunately, will often highlight the need for a relatively large sample size.

3. Comparison against a target

Sample size determination is a context dependent analysis. Thus, the sample size determined for one application may vary in suitability for other, possibly unforeseen, applications. This section focuses on issues connected with the comparison of accuracy, expressed as the proportion of correctly allocated cases, against a target value before turning to issues connected with comparison against another classification in the next section and draws on an established literature on sample size determination (e.g. [2]).

The decision on whether a classification meets the accuracy needs of a user is typically based on the comparison of its accuracy against a threshold target accuracy (e.g. the minimum acceptable accuracy). The target accuracy should be defined for the specific application in-hand, recognising issues such as the nature of the data sets used, classes and intended use of the classification [4]. With a pre-defined target accuracy value, the design of the accuracy assessment programme could be based on basic issues connected with statistical hypothesis testing. The latter, crudely, involves deciding between a null (H_o) and an alternative (H_1) hypothesis. For example, in a standard two-sided test the null hypothesis might be that the two proportions compared did not differ (i.e. a null hypothesis of zero difference) while the alternative would be that they did differ; different scenarios exist but are not considered here. If there was reason to specify a directional component (e.g. that one proportion should be larger than the other) then a one-sided alternative hypothesis could be tested, although it may often be desirable to guard against the unexpected and give results for two-sided testing as well [2]. The decision as to which hypothesis to 'accept' and which to reject would depend on the magnitude of a computed test statistic; although hypotheses are not strictly accepted this expression is used for ease of presentation. The outcome of this type of analysis is prone to two types of error. A Type I error occurs when H_1 is incorrectly accepted. That is, the analysis leads the researcher to incorrectly reject the null hypothesis, of no difference, and declare that a difference exists when in reality it does not. The probability of making a Type I error is typically designated α, the significance level, and is typically set at a low value such as 0.05 (i.e. there is a 5% chance of wrongly calling a difference as being significant). Conversely, a Type II error occurs when the H_o is wrongly accepted and the analyst fails to detect a meaningful difference that actually does exist. The probability of making a Type II error, in which a false H_o is accepted, is typically designated β. Type II errors are conventionally treated as being of lesser importance than Type I errors. Indeed the Type I error is often treated as being about four times more important than the Type II error. So if the 0.05 significance level is adopted, it is common to set β=0.20 [2]. Type II errors can, however, be a major problem as it can often be difficult to interpret non-significant results and these will commonly arise in under-powered studies based on small sample sizes [5].

A fairly typical remote sensing scenario involves the comparison of a classification's output against some other proportion, P_o, that represents the maximum unacceptable accuracy. In this situation, a simple formula may be used to estimate the sample size required to estimate and compare the classification accuracy statements. The aim is to compare the accuracy of a classification, P, against P_o but recognising that there is a minimum detectable effect size, a minimum meaningful difference in accuracy. The researcher needs to define a value of P=P_1 at which the difference between P_1 and P_o (e.g. $|P_1 - P_o|$ for a two-sided test), represents the smallest difference worth detecting. The common remote sensing scenario calls for a one-

sided analysis, with H_o $P \leq P_o$ and H_1 $P > P_o$. If p is the derived estimate of P, the basic test of the difference between proportions, with correction for continuity, is

$$z = \frac{\left|p - P_o\right| - \frac{1}{2n}}{\sqrt{P_o Q_o / n}} \tag{4}$$

where Q_o =1-P_o. At the widely used 95% level of confidence, a difference would be viewed as being statistically significant if $z>|1.96|$. [2] show that once the values for the effect size, α and β have selected the analyst can then derive an initial estimate of the minimum required sample size from

$$n' = \left[\frac{z_\alpha \sqrt{P_0(1-P_0)} + z_\beta \sqrt{P_1(1-P_1)}}{P_1 - P_0} \right]^2 . \tag{5}$$

This may be refined through the inclusion of continuity correction to yield an estimate of the required sample size from

$$n = \frac{n'}{4}\left(1 + \sqrt{1 + \frac{2}{n'\left|P_1 - P_0\right|}}\right)^2 \tag{6}$$

These equations may also be inverted to indicate the power of a test (1-β) based on a specified sample size. Retrospective power analyses can be controversial and, as discussed below, it may be more useful to evaluate confidence intervals fitted to the derived estimates [6]. Power analyses are, however, useful in prospective studies [5]. Power is a function of P but can be estimated for various values of P=P_1. The power of a one-sided test may be derived through the use of equation 7, with the power at P_1 being

$$P\left[Z \leq \frac{\sqrt{n}\left|P_1 - P_0\right| - \frac{1}{2\sqrt{n}} - z_\alpha \sqrt{P_0(1-P_0)}}{\sqrt{P_1(1-P_1)}} \right] \tag{7}$$

For equation 7 to be at least equal to the power (1-β), the expression inside the square bracket must be as large as the critical value z_β. For a two-sided test, $z_{\alpha/2}$ is substituted for z_α [2]. A variety of alternative formulae are given in the literature and which can yield different results and a set of alternative approaches for sample size and power estimation are reviewed in [7]. Critically, however, the required sample size is a direction function of the effect size, α and power (1-β). This has important implications as a very large sample may be required, especially if aiming for high power and a small effect size.

4. Comparing classification outputs

Interest is often focused on differences between classification outputs, especially in the case of the evaluation of different classifiers. The latter is typically based on the comparison of the accuracy with which they classify data. This comparison may be seeking to determine if the accuracy with which data are classified by two classifiers differs significantly. Commonly, the researcher is hoping to identify that a new classifier can be used to derive a more accurate classification than some other, perhaps benchmark, classifier but other scenarios exist. The a researcher might, for example, be seeking to determine if the addition of ancillary data increases accuracy or if a simplified approach or data set, perhaps after feature reduction, maintains the accuracy achieved in a conventional analysis. These aims differ in important ways, some not considered here, but in each the central issue is the comparison of proportions and evaluation of the statistical significance of their difference. For simplicity, this section will assume mainly that the aim is simply to test for the statistical significance of a difference in the proportion of correctly allocated cases (i.e. two-sided test of the difference) but the basis of the discussion may be extended to other scenarios.

The capability of different classifiers is often evaluated on the basis of a comparison of their accuracies. It is inappropriate to simply compare the magnitude of the accuracy measure derived from each classification as these are estimates of classification accuracy and some allowance for the variance of each estimated accuracy measure is required. One approach that has been widely promoted in the remote sensing literature is to base the comparison on the difference in estimated kappa coefficients of agreement determined for the classifications. The kappa coefficient of agreement may be derived from

$$\hat{\kappa} = \frac{p - p_c}{1 - p_c} \tag{8}$$

where p_c is the proportion of cases allocated correctly by chance. The test for a difference between two kappa coefficients, $\hat{\kappa}_1$ and $\hat{\kappa}_2$, is then based on

$$z = \frac{\hat{\kappa}_1 - \hat{\kappa}_2}{\sqrt{\hat{\sigma}^2_{\kappa_1} + \hat{\sigma}^2_{\kappa_2}}} \tag{9}$$

where $\hat{\sigma}^2_{\kappa_1}$ and $\hat{\sigma}^2_{\kappa_2}$ represent the estimated variances of the derived coefficients. A difference is taken to be statistically significant at the 95% level of confidence if z>|1.96|. There are, however, some major concerns with the use of the kappa coefficient as a measure of classification accuracy in remote sensing. In particular, each of the main arguments offered for its use may be flawed and/or apply equally to other measures of accuracy [4]. Given the limitations of the kappa coefficient as a measure of classification accuracy, and recognising that it is no more than a re-scaled version of the proportion of correctly allocated cases, all further discussion is focused on a similar approach formulated for the comparison of proportions.

With accuracy expressed as the proportion of correctly allocated cases, the comparison is based on the difference between proportions. Assuming that the samples used to estimate the proportions are independent of each other, the significance of the difference between two proportions may be estimated from:

$$z = \frac{P_1 - P_2}{\sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \tag{10}$$

where $\bar{p} = \frac{x_1 + x_2}{n_1 + n_2}$ with x_1 and x_2 representing the number of correctly allocated cases in the classifications of data sets of size n_1 and n_2 respectively. With continuity correction this may be expressed as

$$z = \frac{\left|P_1 - P_2\right| - \frac{1}{2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}{\sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \tag{11}$$

Note that this type of analysis is less efficient than the approach used in the evaluation against a target value (equation 4) and so requires a larger sample size [2].

When the aim is to compare proportions to determine if they are different it is first necessary to define what magnitude of difference in the proportions compared is meaningful [2, 7]. This may be difficult to define and may be very application specific. Small differences in accuracy may sometimes be important (e.g. if the costs of misclassification are high). However, a large difference may sometimes be required. For example, a researcher may only wish to adopt the use of a complex classifier if it can be shown that the accuracy derived from its use is very much larger than that from another, easier to use, classifier and so worth the extra effort involved in its use. In order to estimate the sample size required for a comparison the minimum meaningful difference should be defined at the outset. As in the comparison against a target value, this minimum meaningful difference (effect size) is a key component in the planning of sample size determination. Once defined, the aim is to acquire a sample that is large enough to detect the selected effect.

The probability of detecting a specified effect is represented by the power of the test and designated 1-β. In testing for a difference between two proportions (two-sided test), [2] show that with α, 1-β and the effect size selected, an initial estimate of the required sample size from each of the populations being compared may be estimated from

$$n' = \frac{\left(z_{\alpha/2}\sqrt{2\bar{P}\bar{Q}} + z_{\beta}\sqrt{P_1 Q_1 + P_2 Q_2}\right)^2}{\left(P_2 - P_1\right)^2} \tag{12}$$

with $\bar{P} = \frac{P_1 + P_2}{2}$ and $\bar{Q} = 1 - \bar{P}$. Refined with continuity correction the required sample size is

$$n = \frac{n'}{4}\left(1 + \sqrt{1 + \frac{4}{n'|P_2 - P_1|}}\right)^2 \tag{13}$$

The total sample size required would, therefore, be $2n$.

A range of alternative formulae are available for this application as well as for situations such as when the proportions are estimated from samples of unequal size or complex designs [2, 7]. The equations may be inverted to allow estimation of the power of a test using a fixed sample size. Such an assessment is based on

$$z_\beta = \frac{|P_2 - P_1|\sqrt{n - \frac{2}{|P_2 - P_1|}} - z_{\alpha/2}\sqrt{2\bar{P}\bar{Q}}}{\sqrt{P_1Q_1 + P_2Q_2}} \tag{14}$$

that corresponds to the power associated with the specified sample size [2]. Prospective power calculations can play a useful role in helping to design or plan an analysis. It is, however, important to stress that retrospective power analyses can be of limited value.

Often α and β are set at rather arbitrary values following convention [7]. For example, it is common for values such as α=0.05 and β=0.20 to be used. With such values the researcher is accepting a 5% chance of erroneously designating a difference as being significant when in reality it is not and an 80% (1-0.20) chance of finding statistical significance if the specified effect exists. These values may be far from ideal and their use in an analysis can produce unhelpful results. The settings may, for example, result in an under-powered analysis, one that lacks the capability to detect a meaningful difference.

In many studies, the power of the test is not considered explicitly, given the common focus on Type I error, and a value selected arbitrarily. Indeed the researcher may perhaps select a value for the significance level, α, but accept, implicitly by default, a value of β. However, in many studies protection from Type II error, failing to find a difference of meaningful magnitude, may be important and a power of 0.8 inadequate. Increasing the power of the test will, however, normally require an increase in sample size, unless the researcher was, for example, willing to trade Type I and Type II error rates. Similarly, the smaller the effect size specified, the larger the sample size required. The detection of small differences using typical values for α and β may require a large sample. It is, however, important to set values appropriate for the study's aims recognising that a size that is too large or too small can be undesirable.

The discussion has so far has assumed that the proportions being compared are independent. This is often not the case in remote sensing applications, especially those focused on issues such as classifier comparison. Commonly, the same testing set is used in the evaluation of the classifiers being compared. Consequently, the analysis is based upon paired or related samples and not independent samples. While the use of related samples has advantages it means the methods based on independent samples outlined above are no longer appropriate. Instead, the researcher should select an alternative approach that is suited to matched or related samples. In relation to the comparison of proportions, one approach that has been widely used in remote sensing is to use the McNemar test. This is based on a cross-tabulation of the classification outputs (Figure 1) and one popular basis, without continuity correction, is to use

$$z = \frac{f_{10} - f_{01}}{\sqrt{f_{10} + f_{01}}} \tag{15}$$

with f indicating the frequency of cases in a particular element of the matrix. It is evident that the test is based on the discordant cases [2, 8], with an effective sample size of

$$s = f_{10} + f_{01} \tag{16}$$

		Classifier A	
		Correct	Incorrect
Classifier B	Correct	f_{00}	f_{10}
	Incorrect	f_{01}	f_{11}

Figure 1. Summary of the allocations made by 2 classifiers for the McNemar test.

As with tests based on the comparison of proportions discussed above, it is possible to estimate the required sample size for an analysis. This again requires specification of α, β and the effect size. With the effect size denoted δ, [8] argues that the required sample size for a two-sided test may be estimated from a relationship such as

$$n = \frac{\left[z_{\alpha/2}\psi + z_{\beta}[(\psi^2 - \frac{1}{4}\delta^2(3+\Psi)]^{\frac{1}{2}}\right]^2}{\psi\delta^2} \tag{17}$$

where Ψ is the probability of obtaining a discordant pair. For a two-sided test, $z_{\alpha/2}$ may be replaced with z_{α}. A variety of other formulae have been suggested in the literature (e.g. [9]) as well as discussions on the inclusion of continuity correction [2].

The main problem in using equation 17, or one of the alternatives, is that Ψ is unknown before the analysis. However, as there is often some prior knowledge of the proportions to be compared and δ should be specified by the researcher, a conservative estimate of the required sample size could be derived using a value at or near the upper bound of Ψ. Further details including methods for estimation of power, before and after data collection, are given in [8]. Because of concerns with basic hypothesis testing and power analyses it may be more appropriate sometimes to use confidence intervals to help interpret results.

5. Use of confidence intervals

Hypothesis testing provides a basic dichotomous outcome. In many studies, especially those based on small samples, the results may be non-significant and these can be difficult to interpret fully [6, 10]. Retrospective power analysis may show that some studies deriving such non-significant results were under-powered. Many concerns have, however, been raised with retrospective studies, which attempt to estimate power once the data have been acquired, especially if the observed effect size is used [6]. Alternatives are to focus on conditional power or perhaps more generally to use the confidence interval. Indeed the fitting of the latter may greatly aid project planning and the interpretation of results, especially as they summarise the plausible effect sizes supported by the data [6, 10]. For example, [11] show the value of predicted confidence limits. The precision of the estimate can be predicted in advance and, as above, the basic relationships used to determine the required sample size. The basis of this determination is that the difference of interest lies at a distance from the H_o that is a function of the selected α and β as

$$\delta = (z_{\alpha/2} + z_{\beta})\mathrm{SE}_{p_1-p_2} \tag{18}$$

where δ is the difference that can be detected with a power of 1-β and $\mathrm{SE}_{p_1-p_2}$ the standard error of the difference between the proportions. The latter may be written generally as

$$\mathrm{SE}_{p_1-p_2} = \sqrt{\frac{P_1(1-P_1)}{n_1} + \frac{P_2(1-P_2)}{n_2}} \tag{19}$$

The predicted 95% confidence interval is the observed difference $\pm 1.96\,\mathrm{SE}_{p_1-p_2}$. Rewriting equation 18 gives the predicted confidence interval of

$$\text{Observed difference} \pm 1.96\left(\frac{\delta}{z_{\alpha/2} + z_{\beta}}\right). \tag{20}$$

Thus, if using popular values such as α=0.05 and β=0.2 (i.e. power 0.8), the z value for a two-sided test with Type I error of α is $z_{\alpha/2}$ =1.96 and that for a one-sided Type II error β is z_{β} = 0.84. Therefore, the 95% predicted confidence interval would be the observed difference ±0.7δ.

These equations may also be rewritten for sample size determination, yielding a relationship of the form

$$n = \frac{2\overline{P}\overline{Q}(z_{a/2} + z_{\beta})^2}{\delta^2} \tag{21}$$

The above is suitable for comparison of independent proportions. If using related samples and the McNemar test, it is still possible to fit confidence interval to the estimate. For example, one approach discussed by [12] is

$$\text{Observed difference} \pm (z_{\alpha/2}\mathrm{SE}_{\mathrm{M}} + \tfrac{1}{n}) \tag{22}$$

with the standard error of the difference for this test represented by

$$SE_M = \sqrt{\frac{(f_{00} + f_{11})(f_{10} + f_{01}) - 4f_{10}f_{01}}{n^3}} \quad (23)$$

While the use of confidence intervals is focused on estimation rather than hypothesis testing, the two issues are intimately linked. The confidence interval approach could be used as an alternative or accompaniment to other approaches above. The key attraction of the use of a confidence interval is that it provides a simple to derive and easy to interpret summary of what the difference might reasonably be. All the values that lie within the confidence interval are not-refuted by the analysis and so it is possible to gain an appreciation of just how far the population parameter value may deviate from the value specified in the null hypothesis. By indicating all possible differences supported by the data, the confidence interval provides a richer basis for interpretation and allows stronger conclusions to be drawn about the null hypothesis. Critically, rather than just the dichotomous decision from hypothesis testing, which is especially unhelpful the difference is deemed to be not significant, the confidence interval summarises the possible differences supported by the data [10].

6. Conclusions

The size of the testing set has important implications to accuracy assessment and comparison. In particular, differences in accuracy should be tested rigorously. However, like other research communities [6] the power of tests has often not been considered and samples of inappropriate size used. Commonly a relatively small sample size has been used, resulting in tests that only have reasonable power to detect relatively large differences. Such analyses may often yield non-significant results which can be difficult to interpret. With careful planning of the accuracy assessment and comparison programme and use of confidence intervals in the interpretation, however, useful results can be obtained.

Careful planning of an accuracy assessment and comparison activity can help reduce the use of designs that have little chance of producing significant results. Indeed, as resources are often limited and there is a common desire for parsimony, careful planning will also help produce the most appropriate design for the resources available. In relation to sample size required for accuracy evaluation and comparison it is vital to consider fully the goals of the study. The sample size selected should be large enough to provide a sufficient chance of finding a statistically significant difference. However, collecting too much data may be wasteful of resources and has the potential to result in any non-zero difference being declared statistically significant. Perhaps more importantly, a study will have too little power to detect a meaningful or statistically significant result if too few cases are acquired. Using insufficient cases may be a large waste of resources as the study may lack the ability to derive useful results. For the purpose of accuracy comparison, there is a need to specify the sample size for the particular problem at-hand and for this the researcher needs to specify the effect size (the meaningful difference), significance level (α) and power (1-β). Finally, it is suggested here that researchers fit and state confidence intervals to the derived estimates of accuracy as these can greatly enhance the value of the results and aid interpretation.

7. References.

[1] De Leeuw, J., Jia H., Yang, L., Liu, X., Schmidt, K. and Skidmore, A. K., Comparing accuracy assessments to infer superiority of image classification methods, *International Journal of Remote Sensing*, 2006, **27**: 223-232.

[2] Fleiss, J. L., Levin, B. and Paik, M. C., *Statistical Methods for Rates and Proportions*, 3rd edition, Wiley, New Jersey, 2003.

[3] Crews-Meyer, K. A., Hudson, P. F. and Colditz, R. R., Landscape complexity and remote classification in eastern coastal Mexico: applications of Landsat-7 ETM+ data, *Geocarto International*, 2004, **19**(1): 45-56.

[4] Foody, G. M., Harshness in image classification accuracy assessment, *International Journal of Remote Sensing* (in press), 2008.

[5] Colegrave, N. and Ruxton, G. D., Confidence intervals are more useful complement to nonsignificant tests than are power calculations, *Behavioral Ecology*, 2003, **14**: 446-450.

[6] Trout, A. T., Kaufmann, T. J. and Kallmes, D. F., No significant difference… says who? *American Journal of Neuroradiology*, 2007, **28**: 195-197

[7] Sahai, H. and Khurshid, A., Formulae and tables for the determination of sample sizes and power in clinical trials for testing differences in proportions for the two-sample design: a review, *Statistics in Medicine*, 1996, **15**: 1-21.

[8] Miettinen, O. S., The matched pairs design in the case of all-or-none responses, *Biometrics*, 1968, **24**: 339-352.

[9] Connor, R. J., Sample size for testing differences in proportions for the paired-sample design, *Biometrics*, 1987, **43**: 207-211.

[10] Aberson, C., Interpreting null results: improving presentation and conclusions with confidence intervals, *Journal of Articles in Support of the Null Hypothesis*, 2002, **1**: 36-42.

[11] Goodman, S. N. and Berlin, J. A., The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results, *Annals of Internal Medicine*, 1994, **121**: 200-206.

[12] Newcombe, R. G., Improved confidence intervals for the difference between binomial proportions based on paired data, *Statistics in Medicine*, 1998, **17**: 2635-2650.