Classification accuracy comparison: hypothesis tests and the use of confidence intervals in evaluations of difference, equivalence and non-inferiority.

Giles M. Foody
School of Geography
University of Nottingham
Nottingham
NG7 2RD
UK

Abstract

The comparison of classification accuracy statements has generally been based upon tests of difference or inequality when other scenarios and approaches may be more appropriate. Procedures for evaluating two scenarios with interest focused on the similarity in accuracy values, non-inferiority and equivalence, are outlined following a discussion of tests of difference (inequality). It is also suggested that the confidence interval of the difference in classification accuracy may be used as well as or instead of conventional hypothesis testing to reveal more information about the disparity in the classification accuracy values compared.

1. Introduction

Classification accuracy statements are often compared in remote sensing research. Such comparisons often form the basis of studies that have sought, amongst other things, to evaluate image pre-processing methods (Song *et al*., 2001), classifiers (Mitrakis *et al*., 2008; Ngigi *et al*., 2008) and impacts of sensor properties such as spatial and spectral resolution (Gao and Liu, 2008) on the ability to discriminate classes. The central focus of such comparative analyses has been the magnitude of the difference in the accuracy values contained in the classification accuracy statements.

Quantitative comparisons of accuracy have typically been achieved using popular hypothesis testing approaches based on tests of the statistical significance of the difference or inequality in the values observed. Studies have generally used either a popular approach for testing the statistical significance of a difference, such as the comparison of kappa coefficients (Congalton *et al*., 1983; Sha *et al*., 2008) or proportion of correctly allocated cases (Gao and Liu, 2008), or recently promoted approaches such as the McNemar test (Foody, 2004; De Leeuw *et al*., 2006; Demir and Ertürk, 2008). Most studies have focused on the difference in accuracy, irrespective of the direction of the difference, although one-sided hypothesis testing may be used to help studies with a directional component. Although popular, there are, however, sometimes problems with the use of statistical tests of difference (inequality). For example, the use of an inappropriate sample size of testing cases can be a major problem (Foody, 2009). If the sample size is too large, a comparative study may ascribe statistical significance to any non-zero difference in accuracy, even if the difference is trivially small and so not of a meaningful or important magnitude. Conversely, if the sample size is too small, the comparative evaluation may be lacking in power and so may not result in the detection of a large and meaningful difference. Additionally, the focus on the straightforward difference or inequality in accuracy values may not always be appropriate and the limitations of hypothesis testing based approaches should be recognised.

Although comparative tests focus on the difference in the magnitude of the accuracy values, the nature of the difference and test may vary greatly. In distinguishing between types of statistical test, the main concern in this paper is on the objective of study and hence the hypotheses evaluated. With the popular test for a difference (inequality), for example, the alternative to the null hypothesis evaluated is that the accuracy values are unequal or different. Other types of test, albeit based on differences in the magnitude of accuracy values, evaluate different hypotheses as a function of study objectives. Sometimes the objective of a study is not really focused on the detection of a difference (inequality) in accuracy but on an evaluation of the similarity in accuracy, a subtle but important distinction. The use of a statistical test of the difference (inequality) in such a study may have some value but would often be inappropriate. The use of such a test could, for example, show that a statistically significant difference existed, and so might be used to suggest that the accuracy statements that have been compared could not be considered similar. Alternatively, the test may lead to the conclusion that no significant difference exists. This result may, however, simply be a consequence of the test lacking sufficient power due to the use of a small sample (Hoenig and Heisey, 2001; Aberson, 2002; Trout, 2007). Perhaps more importantly, non-significant results present a philosophical concern, with the failure to find a difference not being proof of similarity (Altman and Bland, 1995; Barker *et al*., 2002; Carlin and Doyle, 2002; Martinez-Abrain, 2007). This is a major concern for studies that seek to evaluate similarity in accuracy values.

This brief paper aims to illustrate the statistical basis for a variety of comparative analyses. Specifically, the article addresses some issues in the statistical testing for three types of comparison. For clarity, tests will be named throughout in relation to the alternative hypothesis evaluated in the comparison. The discussion will show how hypothesis testing approaches may be used for each type of comparative analysis but, in response to limitations, will also promote the use of confidence intervals as a general basis for the comparison of accuracy values. The article draws on an extensive literature in other disciplines in the hope that it may be a step in the advancement of methodological practice in remote sensing.

2. Comparative tests

This section will consider three types of comparative statistical testing: tests of difference (inequality), equivalence and non-inferiority. All three tests focus on the difference in the magnitude of the accuracy values compared but evaluate a dissimilar set of hypotheses. The discussion will first illustrate how each type of comparison may be undertaken within a conventional hypothesis testing framework before considering the potential of confidence intervals as a basic tool in classification comparison. Other types of comparative analyses, such as non-superiority testing, which may be of only rare use in remote sensing and which are generalisations of the scenarios outlined will not be discussed. Additionally, while sampling issues are important (Stehman, 1997; 2000) these will not be considered further so that the focus may be on the general nature of comparative analyses. It is, however, assumed that the samples used were acquired by simple random sampling and that the estimates derived are unbiased.

Statistical tests of the difference (inequality) in accuracy values are commonly encountered in remote sensing. For example, many studies have sought to evaluate a set of classifiers and have done so on the basis of the accuracy with which they can classify data (De Leeuw *et al*., 2006). Using this common application as a basis for discussion, this section outlines key issues in statistical comparisons. These typically involve comparing the accuracy of, say, a new classifier against that derived from the application of a standard classifier. Thus, for instance, one of the most widely promoted means to compare classification accuracy statements in remote sensing is through the comparison of kappa coefficients. With this approach, the statistical significance of the difference between two independent kappa coefficients is evaluated through the calculation of a *z* value with

$$z = \frac{\hat{\kappa}_1 - \hat{\kappa}_0}{\sqrt{\hat{\sigma}^2_{\kappa_1} + \hat{\sigma}^2_{\kappa_0}}} \qquad (1)$$

where $\hat{\kappa}_0$ and $\hat{\kappa}_1$represent respectively the estimated kappa coefficients for the classifications derived from the standard and new classifiers and $\hat{\sigma}^2_{\kappa_0}$ and $\hat{\sigma}^2_{\kappa_1}$ their associated variances (Congalton *et al*., 1983; Rosenfield and Fitzpatrick-Lins, 1986). Assuming a normal distribution, a difference is taken to be statistically significant if $|z|>z_{\alpha/2}$ where $z_{\alpha/2}$ is taken to represent the value cutting off the proportion α/2 in the standard normal curve's upper tail and may be determined from statistical tables. Thus, if

α=0.05 and equation 1 yielded z>1.96 or z<-1.96 the difference would be declared significant at the 5% significance level. The approach may also be adapted for the situation in which related rather than independent samples have been used (McKenzie *et al*., 1996; Donner *et al*., 2000; Foody, 2004). There are, however, many concerns about the use and interpretation of the kappa coefficient (Stehman, 1997; Pontius and Millones, 2008; Foody, 2008) which may make it preferable to sometimes express accuracy as the proportion of correctly allocated cases and base the comparison on a test of the difference between proportions.

When comparing proportions, the aim is generally to make an inference about the population proportions, P_0 and P_1, from the proportions estimated from samples, p_0 and p_1. The statistical significance of a difference between two proportions may be evaluated through

$$z = \frac{|p_1 - p_0|}{\sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n_1} + \frac{1}{n_0}\right)}} \qquad (2)$$

where $\bar{p} = \frac{x_1 + x_0}{n_1 + n_0}$ with x_0 and x_1 representing the number of correctly allocated cases in the classifications of samples of size n_0 and n_1 respectively (Fleiss *et al*., 2003). It may sometimes be appropriate to apply a correction for continuity. This involves making a minor adjustment to the numerator of equation 2 with the test then based on

$$z = \frac{|p_1 - p_0| - \frac{1}{2}(\frac{1}{n_1} + \frac{1}{n_0})}{\sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n_1} + \frac{1}{n_0}\right)}} \qquad (3)$$

Continuity correction only has a major impact on tests when the sample sizes are small and is not considered further in this paper. Further discussion on continuity correction, including formulae for a variety of scenarios, is given by Fleiss *et al*. (2003).

Comparative studies founded on equations 1-3 (or similar) are based on conventional statistical hypothesis testing. In this, two competing hypotheses are evaluated. These are the null hypothesis (H_o) which states that there is no difference in accuracy (i.e. P_0-P_1=0 or P_0=P_1) and an alternative hypothesis (H_1) that the accuracy values differ (i.e. $P_1 \neq P_0$); a directional hypothesis may sometimes be suitable and stated. The derived *z* value is interpreted in the same way as for the comparison of kappa coefficients, with the null hypothesis rejected and a difference viewed as being statistically significant if $|z| > z_{\alpha/2}$. Within a deductive scientific methodology based on the principle of falsification, it is the rejection of the null hypothesis that is particularly useful, showing that the difference is not the value stated in the null hypothesis. Rejection of the null hypothesis, of no difference in accuracy, is central to the declaration that a difference exists.

A popular refinement of the above approaches is to report also the '*p*-value' for the outcome of a test. The latter is the probability of obtaining the derived test statistic or one more extreme if the null hypothsis is true; the '*p*-value' is a standard expression and should not be confused with symbols used in this paper to indicate proportions. As the magnitude of the '*p*-value' is directly related to the plausibility of the null hypothesis (Carlin and Doyle, 2002), it provides more detail on the outcome of a test than a dichotomous decision based on a single threshold level of probability.

In planning a study that involves a test of the significance of a difference in proportions the analyst should define three attributes or conditions. These conditions are the significance level (α), the power of the test (1-β) and the minimum meaningful difference in accuracy or effect size. The α and β define the likelihood of making a type I error (rejecting H_o when it is true) and a type II error (failing to reject H_o when it is false) respectively. The effect size should be specified by the analyst for the task at-hand. With these three conditions defined at the beginning of an analysis it may be possible to define the required sample size for the construction of a testing set to meet the objectives of a study (Fleiss *et al*., 2003; Foody, 2009). While such an approach may often be useful and help ensure that a study is designed to meet its objectives it must be noted that sometimes studies are not focused on determining if a difference in accuracy exists but whether one is, essentially, absent or insignificantly small. While the remainder of this section is focused on tests with an alternative aim to difference (inequality) testing, the design considerations still require attention.

Often the desire in remote sensing studies is not to test for a difference (inequality) in accuracy but to evaluate the similarity in accuracy values. Two scenarios are commonly encountered. First, some studies aim to assess non-inferiority. That is, the study might seek to show that one classification is not substantially worse, in terms of accuracy, than another. For example, the aim might be to determine if a new method yields a classification that is not substantially worse than that from a standard method. This might be the case when some simplifying approach or approximation is under evaluation. For example, in testing a fast approximation of the maximum likelihood classification (Settle and Briggs, 1987) or a knowledge-based procedure for training site selection (Foody and Mathur, 2004; Mathur and Foody, 2008) or perhaps in some feature reduction analyses (Demir and Ertürk, 2008; Wang and Li, 2008) the aim is commonly to maintain the level of accuracy derived from some standard approach. The second scenario of testing in relation to evaluations of similarity relates to studies that aim to evaluate equivalence. That is, a study that seeks to show that the two classifications are comparable or similar in terms of their accuracy. This might be the case when the aim is to compare classifications derived using imagery acquired by different versions of the same sensing system. Given concerns on sensor data continuity (Wulder *et al*., 2008) one might wish to ensure that classifications of a site derived from old and new versions of a sensor are of comparable accuracy. In such a situation, the aim is to show the accuracy values are similar. That is, the accuracy values derived may differ slightly, and in either direction, but not by an appreciable or important amount. Both non-inferiority and equivalence may be evaluated using a standard hypothesis testing approach similar to that used in testing for a difference (inequality) in accuracy. This can be readily illustrated for comparative analyses of proportions.

Key features of both of the non-inferiority and equivalence testing scenarios are the specification of a degree of indifference, a tolerable level of a difference in accuracy, and a change in the hypotheses tested. In relation to the latter, it is important to note the change in emphasis given the desire to work within a scientific paradigm based on the principle of falsification. With non-inferiority and equivalence testing there is a transposition of what might normally be expected to be the null and alternative hypotheses (Fleiss *et al*., 2003). For example, with equivalence testing, the burden of proof is reversed from the normal situation with testing for a difference and recognises that it is insufficient to fail to show a difference as the researcher must be highly confident that a large difference does not exist (Hoenig and Heissey, 2001). Critically, the null hypothesis stated is no longer one of no difference but is instead one of a large difference. Thus the null hypothesis is, essentially, that the accuracy values differ by a magnitude larger than a tolerably small amount while the alternative hypothesis is one of similarity (Barker *et al*., 2002; Colegrave and Ruxton, 2003). The resulting test is one that can provide proof of similarity. Although researchers are often conditioned to use a null hypothesis of no difference it is essential to note that there is choice over the value stated in the null hypothesis and this is central to both non-inferiority and equivalence testing (Hoenig and Heisey, 2001). Again it is important to remember that useful scientific information is gained through the rejection of the null hypothesis and so a null hypothesis of no difference would be unhelpful in some studies.

For a test of non-inferiority, the researcher needs to specify the limit of indifference. The latter is, essentially, a measure of largest reduction in accuracy that could be considered to be unimportant. With this limit of indifference defined as Δ, if the two classifiers differ in accuracy by a value < Δ then non-inferiority may be inferred. More formally, if P_o is the proportion of correctly allocated cases observed from the standard method and P_1 the proportion correct derived from the classifier being evaluated, the null hypothesis of the test, H_o, is $P_1 \leq P_o - \Delta$ (i.e. that P_1 is inferior) and H_1 is $P_1 > P_o - \Delta$ (i.e. that that P_1 is non-inferior). Fleiss *et al.* (2003) show that the test itself is then based on the critical region $z \geq z_\alpha$, where

$$z = \frac{p_1 - p_0 + \Delta}{\sqrt{\frac{\hat{P}_0 \hat{Q}_0}{n_0} + \frac{(\hat{P}_0 - \Delta)(\hat{Q}_0 + \Delta)}{n_1}}} \tag{4}$$

where $\hat{P}_0 = 1 - \hat{Q}_0$ and is the maximum likelihood estimate of the standard classifiers success rate when the difference between proportions is at the limit of indifference and satisfies

$$\frac{p_0 - \hat{P}_0}{\hat{P}_0 (1 - \hat{P}_0) / n_0} + \frac{p_1 - (\hat{P}_0 - \Delta)}{(\hat{P}_0 - \Delta)(1 - \hat{P}_0 + \Delta) / n_1} = 0 \tag{5}$$

Solving equation 5 for $\hat{P}_0$ allows the test based on equation 4 to be undertaken. Rejection of the null hypothesis provides evidence in support of the alternative hypothesis of non-inferiority.

With a test of equivalence, a zone of indifference is specified around P_o. Should accuracy values differ by an amount that lies within this zone they would be deemed to be of equivalent accuracy. The null hypothesis is that the accuracy values are not equivalent (i.e. that the accuracy values differ). The desire in an equivalence trial is, essentially, to reject this hypothesis in favour of the alternative hypothesis of equivalence. More formally, and assuming for simplicity a symmetrical zone of indifference $|P_1\text{-}P_o|<\Delta$, the hypotheses may be considered as a H_o of $P_1 \leq P_o\text{-}\Delta$ or $P_1 \geq P_o+\Delta$ and a H_1 of $|P_1\text{-}P_o|<\Delta$. The H_o would be rejected when both one-sided components would be rejected. Fleiss *et al*. (2003) show that in this situation H_o is rejected when $z' \geq z_\alpha$ and $z'' \leq -z_\alpha$ where z' is given by equation 4 and z'' is from

$$z'' = \frac{p_1 - p_0 - \Delta}{\sqrt{\frac{(\hat{P}_1 - \Delta)(\hat{Q}_1 + \Delta)}{n_0} + \frac{\hat{P}_1\hat{Q}_1}{n_1}}} \tag{6}$$

where $\hat{P}_1$ represents the maximum likelihood estimate for the proportion correct from classifier 1 when the difference between proportions is at the upper limit of indifference with $P_1\text{-}P_0=\Delta$ and satisfies

$$\frac{p_0 - (\hat{P}_1 - \Delta)}{(\hat{P}_1 - \Delta)(1 - \hat{P}_1 + \Delta)/n_0} + \frac{p_1 - \hat{P}_1}{\hat{P}_1(1 - \hat{P}_1)/n_1} = 0 \tag{7}$$

Solving equation 7 for $\hat{P}_1$ allows the test based on equation 6 to be undertaken

As before, it is the rejection of the null hypothesis that provides useful information, providing evidence in support of the view that the accuracy values are equivalent. As with tests of difference (inequality), both non-inferiority and equivalence testing are sensitive to sample size issues. An equivalence test based on a very small sample may, for example, commonly be expected to suggest equivalence but the result is of little value as the test is lacking in power (Tai and Lee, 1998). To help plan accuracy assessment programmes, equations for sample size selection may be used (e.g. Fleiss *et al*., 2003).

Although widely used, the comparison of proportions by methods such as those outlined above is often only appropriate for use with data sets that may be considered to be independent. This is often not the case in remote sensing applications. Commonly, for example, the same testing set is used in comparative studies. Thus, for example, in studies evaluating a set of classifiers it is common for the same testing set to be used to aid like-for-like comparison. In such circumstances the assumption of independence that underlies the methods described above is unsatisfied and an alternative techniques should be used (Foody, 2004; De Leeuw, 2006). One such technique that may sometimes be appropriate as a test for the difference (inequality) of proportions is the McNemar test. Although the latter has been adopted in remote sensing studies as a tool for evaluating the significance of a difference (inequality) in accuracy when a common testing set has been used it, or closely related techniques, may also be used for tests of equivalence and non-inferiority (Nam, 1997; Newcombe, 1998; Tango, 1998). However, rather than seek an alternative statistical test to use within a hypothesis testing framework it may be preferable to consider the use of an alternative approach for accuracy comparison which is of broader applicability and conveys additional information. One attractive approach is to base comparative analyses on the confidence interval fitted to the estimated parameter. The latter has the potential to form a more general basis for classification accuracy comparison, whether in testing for difference (inequality), non-inferiority or equivalence.

3. Confidence intervals

One problem with the conventional hypothesis testing process is that it provides a basic dichotomous outcome in which the null hypothesis is either rejected or not. Although further information may be conveyed by the provision of a '*p*-value' this is not always useful and is often mis-interpreted (Goodman, 1999; Di Stefano, 2004; Morgan *et al*., 2005). One major concern, for example, is that a '*p*-value' may highlight statistical significance but this need not relate to practical significance (Di Stefano, 2004; Morgan *et al*., 2005; Kay, 2007). Furthermore while rejecting a null hypothesis is scientifically useful and indicates that the value stated in the null hypothesis is unlikely to be observed in the population it gives no indication of the likely magnitude of the difference in accuracy that exists. That is, the '*p*-value' conveys no information on the possible size of difference that may occur. Thus, for example, if a difference is determined to be significant the hypothesis testing approach only indicates what the effect size may not be (i.e. the difference is not the value in the null hypothesis) rather than what it may be. Interpretation is also difficult if the result is non-significant and may simply indicate that a study was underpowered (Hoenig and Heissey, 2001; Aberson, 2002; Trout, 2007; Foody, 2009). Since the '*p*-value' indicates the probability of obtaining a result as or more extreme than the one obtained, under the assumption of the null hypothesis, it relates to the risk of making a type I error (Goodman, 1999; Trout *et al*., 2007). The '*p*-value' does not provide information in relation to the risk of making a type II error which is a function of the power of the test undertaken. A more useful approach may be to base the comparative assessment of accuracies, whether seeking to evaluate non-inferiority, equivalence or difference (inequality), on confidence intervals fitted to the estimated difference. Confidence intervals may also be used in relation to a variety of scenarios and for results derived from use of both independent and related samples.

The confidence interval provides a range of values within which the population parameter is likely to lie. Assuming a normal distribution, the general expression of the confidence interval is

$$\text{Estimate} \pm z_{\alpha/2}(\text{SE}) \qquad (8)$$

where SE is the standard error of the estimate and, if α=0.05, z=1.96. The provision of confidence limits to accompany classification accuracy statements has been encouraged (Stehman, 1997, 2000; Foody, 2008) but they may be particularly useful in comparative analyses. With the latter interest is focused on the confidence interval of the difference (inequality) in accuracy values. The confidence interval of the difference between two proportions is

$$p_1 - p_0 \pm z_{\alpha/2}\,\text{SE}_{p_1 - p_0} \qquad (9)$$

where $\mathrm{SE}_{p_1-p_0}$ is the standard error of the difference between the estimated proportions. The latter may be viewed generally as

$$\mathrm{SE}_{p_1-p_0} = \sqrt{\mathrm{SE}_1^2 + \mathrm{SE}_0^2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_0(1-p_0)}{n_0}} \qquad (10)$$

Note that the standard error of the difference in equation 10 differs from that represented in the denominator of a standard test for the difference between two proportions, such as that based on equation 2, since an assumption of common population proportions is made in the latter.

The confidence interval derived from equation 9 is appropriate when the samples used were independent and could usefully accompany the outcome of a conventional hypothesis based test. If related samples have been used an alternative approach based on the McNemar test may be more appropriate. A confidence interval for the difference between related proportions to accompany the results of a McNemar test may be achieved following equation 8 using an appropriate estimate of the standard error, SE_M (Lloyd, 1990; Newcombe, 1998; Tango, 1998; 1999). For example, one method for the calculation of the standard error is

$$\mathrm{SE}_\mathrm{M} = \frac{\sqrt{b+c-(b-c)^2/n}}{n} \qquad (11)$$

where *b* and *c* represent the frequency of the discordant pairs (the cases for which the two classifiers compared differed in labelling, which lie in the off-diagonal elements of the 2x2 confusion matrix used in the McNemar test) in the sample of size *n* (Newcombe, 1998); although there are concerns with methods of estimating the confidence interval for use in association with a McNemar test (Lloyd, 1990; Newcombe, 1998). Confidence intervals may also be fitted for use in relation to other tests used in comparing classifications derived from remotely sensed data such as the receiver operating characteristics (ROC) curve (Kerekes, 2008).

To facilitate appreciation of the value of confidence intervals in comparative analyses Figure 1 provides some illustrative scenarios in relation to difference (inequality) and equivalence testing. For scenarios I and II, the entire confidence interval lies outside the zone of indifference and, consequently, also does not span 0. In each of these cases, therefore, the analyst has evidence that the difference is statistically significant as well as evidence that the classifications evaluated are not equivalent. In addition, the confidence intervals give a guide to the range of possible differences that might be expected to occur and highlight the greater uncertainty of the estimated difference in scenario II as opposed to scenario I. With scenario III the confidence interval lies entirely within the zone of indifference providing evidence of equivalence. In addition, the confidence interval includes 0 and so gives evidence that suggests that the classifications compared do not differ significantly in accuracy. This differs slightly from scenario IV in which there is evidence of a statistically significant difference but also that the magnitude of this difference is small, lying within the zone of indifference and so providing evidence for equivalence within the definition of indifference adopted. Finally, scenarios V and VI provide somewhat ambiguous evidence, especially with regard to equivalence. With scenario V there is evidence that a statistically significant difference exists, as the confidence interval does not cross 0, but part of it lies within the zone of indifference. With scenario VI the confidence interval is wide, with part lying inside the zone of indifference and spanning 0, and indicates that further study, perhaps with a larger sample, may be required to evaluate the disparity in accuracy values more fully. In each scenario, the confidence interval may be used to determine whether the null hypothesis specified in a test is rejected or not but also conveys information on the likely range of values in which the difference lies. Additionally, the definition of the zone of indifference is useful in providing context to an interpretation. For example, while the difference observed in scenario IV is statistically significant it is of no practical significance as less than Δ.

The use of confidence intervals does not remove the problems associated with sample size and power that impact on hypothesis tests. For example, the width of the confidence interval will be large if the sample size is small. Indeed, confidence intervals and hypothesis testing are intimately linked (Daly, 1991; Thomas, 1997). The confidence interval may, however, provide a richer assessment to use either independently or in addition to the hypothesis testing based approaches. The confidence interval provides a simple summary of what the difference in classification accuracy might reasonably be. All of the values that lie within the confidence interval derived in an analysis are, for example, not-refuted by the analysis. It is, therefore, possible to appreciate how far the population parameter value may deviate from the value specified in the null hypothesis (Hoenig and Heisey, 2001; Aberson, 2002). If, for example, the 95% confidence interval does not include the value identified in the null hypothesis (0 in the popular situation of difference (inequality) testing), the null hypothesis can be rejected at the 0.05 level of significance and its width illustrates the range of plausible effects sizes or differences that exists (Daly, 1991; Di Stefano, 2004). Since standard hypothesis testing does not indicate the size of the difference, considerable additional information is, therefore, conveyed to the analyst by the confidence interval.

In conclusion, therefore, the common objectives for the comparison of accuracy values encountered in remote sensing may be met through the application of tests of difference (inequality) and similarity. While hypothesis test based approaches can be undertaken for each of the three scenarios of test considered in this paper and has attractions (e.g. ease of planning for sample size to satisfy project objectives) there are concerns which may be addressed through the use of confidence intervals. The latter can be used to derive the same dichotomous decision as a basic hypothesis test but also give valuable information on plausible effect sizes. The use of confidence intervals may also help differentiate statistical and practical significance. In many situations it may be useful to view confidence intervals as adding information to the outcome of a hypothesis test and so provision of both may be helpful to interpreting the results of a comparative study (Di Stefano, 2004; Kay, 2007; Trout *et al*., 2007).

4. Summary and conclusions

Classification accuracy statements are commonly compared in remote sensing studies. Quantitative evaluations undertaken have generally focused on the statistical significance of the difference or inequality in accuracy values. There are, however, many scenarios in which this type of testing is inappropriate. In particular, with studies that are addressing similarities in accuracy it may be more appropriate to test for non-inferiority or for equivalence depending on the specific nature of the study. A key feature of these latter types of test is the use of a null hypothesis of the existence of a difference (as opposed to the widely used null hypothesis of no difference that is the basis of difference testing) which allows the test to provide evidence of similarity in accuracy values. One concern with all of the basic hypothesis testing approaches, however, is that they give only a binary output, indicating if the null hypothesis is rejected or not. Although this is often very useful, the testing does not convey any information on what the magnitude of the difference in accuracy may be, it only indicates what it is unlikely to be. A richer basis for classification comparisons may be provided by the interpretation of confidence intervals. The confidence interval provides a summary of the plausible sizes of the difference in accuracy that are supported by the data (Aberson, 2002; Colegrave and Ruxton, 2003; Trout *et al*., 2007). Thus the confidence interval provides a richer basis for interpretation and allows stronger conclusions to be drawn about the null hypothesis than standard hypothesis testing (Aberson, 2002; Di Stefano, 2004; Martinez-Abrain, 2007). Indeed, the confidence interval provides information well beyond the basic dichotomous decision from hypothesis testing, which may be helpful if the difference is significant or not.

Acknowledgements

I am very grateful to the three referees for their detailed and highly constructive comments which greatly enhanced this paper.

References

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

Altman, D. G. and Bland, J. M. (1995) Absence of evidence is not evidence of absence, *British Medical Journal*, 311, 485.

Barker, L. E., Luman, E. T., McCauley, M. M. and Chu, S. Y. (2002) Assessing equivalence: an alternative to the use of difference tests for measuring disparities in vaccination coverage, *American Journal of Epidemiology*, 156, 1056-1061.

Carlin, J. B. and Doyle, L. W. (2002) Statistics for clinicians, *Journal of Paediatrics and Child Health*, 38, 300-304.

Congalton, R. G., Oderwald, R. G. and Mead, R. A. (1983) Assessing Landsat classification accuracy using discrete multivariate analysis statistical techniques, *Photogrammetric Engineering and Remote Sensing*, 49, 1671-1678.

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

Daly, L. E. (1991) Confidence intervals and sample sizes: don't throw out all your old sample size tables, *British Medical Journal*, 302, 333-336.

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

Demir, B. and Ertürk, S. (2008) Phase correlation based redundancy removal in feature weighting band selection for hyperspectral images, *International Journal of Remote Sensing*, 29, 1801-1807.

Di Stefano, J. (2004) A confidence interval approach to data analysis, *Forest Ecology and Management*, 187, 173-183.

Donner, A., Shoukri, M. M., Klar, N. and Bartfay, E. (2000) Testing the equality of two dependent kappa statistics, *Statistics in Medicine*, 19, 393-387.

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

Foody, G. M. (2004) Thematic map comparison: evaluating the statistical significance of differences in classification accuracy, *Photogrammetric Engineering and Remote Sensing*, 70, 627-633.

Foody, G. M. (2008) Harshness in image classification accuracy assessment, *International Journal of Remote Sensing*, 29, 3137-3158.

Foody, G. M. (2009) Sample size determination for image classification accuracy assessment and comparison, *International Journal of Remote Sensing*, (in press).

Foody, G. M. and Mathur, A. (2004) Toward intelligent training of supervised image classifications: directing training data acquisition for SVM classification, *Remote Sensing of Environment*, 93, 107-117.

Gao, J. and Liu, Y. (2008) Mapping land degradation from space: a comparative study of Landsat ETM+ and ASTER data, *International Journal of Remote Sensing*, 29, 4029-4043.

Goodman, S. N. (1999) Toward evidence-based medical statistics, 1: The *P* value fallacy, *Annals of Internal Medicine*, 130, 995-1004.

Hoenig, J. M. and Heisey, D. M. (2001) The abuse of power: the pervasive fallacy of power calculations for data analysis, *The American Statistician*, 55, 1-6.

Kay, R. (2007) *Statistical Thinking for Non-Statisticians in Drug Regulation*, Wiley, Chichester.

Kerekes, J. (2008) Receiver operating characteristic curve confidence intervals and regions, *IEEE Geoscience and Remote Sensing Letters*, 5, 251-255.

Lloyd, C. D. (1990) Confidence intervals from the difference between two correlated proportions, *Journal of the American Statistical Association*, 85, 1154-1158.

Martinez-Abrain, A. (2007) Are there any differences? A non-sensical question in ecology, *Acta Oecologica*, 32, 203-206.

Mathur, A. and Foody, G. M. (2008) Crop classification by support vector machine with intelligently selected training data for an operational application, *International Journal of Remote Sensing*, 29, 2227-2240.

McKenzie, D. P., Mackinnon, A. J., Peladeau, N., Onghena, P., Bruce, P. C., Clarke, D. M., Haarrigan, S. and McGorry, P. D. (1996) Comparing correlated kappas by resampling: is one level of agreement significantly different from another?, *Journal of Psychiatric Research*, 30, 483-492.

Mitrakis, N. E., Topalogou, C. A., Alexandridis, T. K., Theocharis, J. B. and Zalidis, G. C. (2008) A novel self-organising neuro-fuzzy multilayered classifier for land cover classification of a VHR image, *International Journal of Remote Sensing*, 29, 4061-4087.

Morgan, G., Vaske, J. J. and Harmon, R. J. (2005) *Understanding and Evaluating Research in Applied and Clinical Settings*, Lawrence Erlbaum Associates, Mahwah NJ.

Nam, J. (1997) Establishing equivalence of two treatments and sample size requirements in matched-pairs design, *Biometrics*, 53, 1422-1430.

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

Ngigi, T. G., Tateishi, R., Shalaby, A., Soliman, N. and Ghar, M. (2008) Comparison of a new classifier, the mix-unmix classifier, with conventional hard and soft classifiers, *International Journal of Remote Sensing*, 29, 4111-4128.

Pontius, R. G. and Millones, M. (2008) Problems and solutions for kappa-based indices of agreement, Paper presented at the International Conference on Studying, Modeling and Sense Making of Planet Earth, 1-6 June, Lesvos, Greece, 8pp.

Rosenfield, G. H. and Fitzpatrick-Lins, K. (1986) A coefficient of agreement as a measure of thematic classification accuracy, *Photogrammetric Engineering and Remote Sensing*, 52, 223-227.

Settle, J. J. and Briggs, S. A. (1987) Fast maximum likelihood classification of remotely-sensed imagery, *International Journal of Remote Sensing*, 8, 723-734.

Sha, Z., Bai, Y., Xie, Y., Yu, M. and Zhang, L. (2008) Using a hybrid fuzzy classifier (HFC) to map typical grassland vegetation in Xilin River Basin, Inner Mongolia, China, *International Journal of Remote Sensing*, 29, 2317-2337.

Song, C., Woodcock, C. E., Seto, K. C., Lenney, M. P. and Macomber, S. A., 2001, Classification and change detection using Landsat TM data: when and how to correct atmospheric effects? *Remote Sensing of Environment*, 75, 230-244.

Stehman, S. V. (1997) Selecting and interpreting measures of thematic classification accuracy, *Remote Sensing of Environment*, 62, 77-89.

Stehman, S. V. (2000) Practical implications of design-based sampling inference for thematic map accuracy assessment, *Remote Sensing of Environment*, 72, 35-45.

Tai, B-C. and Lee, J. (1998) Sample size and power calculations for comparing two independent proportions in a 'negative' trial, *Psychiatry Research*, 80, 197-200.

Tango, T. (1998) Equivalence test and confidence interval for the difference in proportions for the paired-sample design, *Statistics in Medicine*, 17, 891-908.

Tango, T. (1999) Improved confidence intervals for the difference between binomial proportions based on paired data, *Statistics in Medicine*, 18, 3511-3513.

Thomas, L. (1997) Retrospective power analysis, *Conservation Biology*, 11, 276-280.

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

Wang, Y. Y. and Li, J., (2008) Feature-selection ability of the decision-tree algorithm and the impact of feature-selection/extraction on decision-tree results based on hyperspectral data, *International Journal of Remote Sensing*, 29, 2993-3010.

Wulder, M. A., White, J. C., Goward, S. N., Masek, J. G., Irons, J. R., Herold, M., Cohen, W. B., Loveland, T. R. and Woodcock, C. E. (2008) Landsat continuity: issues and opportunities for land cover monitoring, *Remote Sensing of Environment*, 112, 955-969.

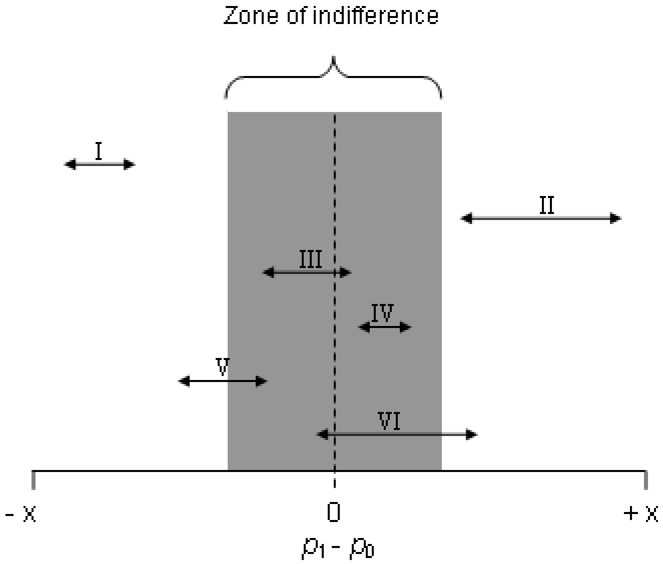

Figure 1. Six scenarios to illustrate the interpretation of confidence intervals in relation to difference and equivalence testing based on the comparison of proportions. The horizontal axis shows the computed difference between two proportions on an arbitrary scale (-x, +x) centered on the point of no difference. For each scenario shown, the width of the arrow depicts the confidence interval at a desired level of significance (e.g. 95% level) which is centred on the observed difference in accuracy. Additionally, with each scenario it has been assumed that the same zone of indifference, highlighted in grey, applies.