Accuracy assessment methods and challenges

Giles M. Foody
School of Geography
University of Nottingham

giles.foody@nottingham.ac.uk

Background

Need for accuracy assessment established.

Considerable progress – now see use of probability sampling, provision of confidence intervals/SE *etc*.

BUT, challenges remain as major errors, biases and uncertainties remain.

Challenges include

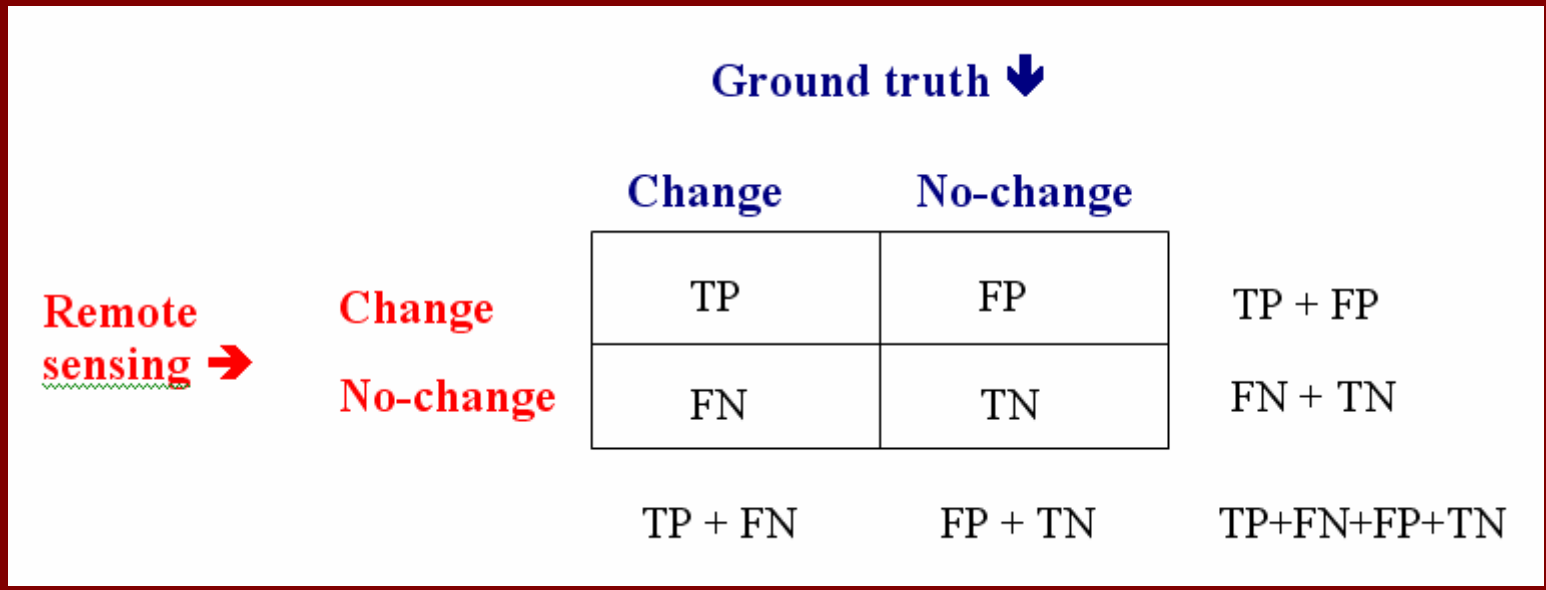
- Class definition (what is a forest?)
- Definition of change – modification *v* conversion *etc*.
- Impacts of spatial mis-registration
- Inter-senor radiometric calibration
- Variations in sensor properties (spatial resolution *etc*)
- Impacts of time of image acquisition
- Required precision of estimation
- Rarity and sampling issues

etc. etc.

Here focus on 2 issues connected with the ground reference data – ‘quality and size’

PART 1: Error matrix Interpretation

Commonly evaluate accuracy with basic binary confusion matrix



		Ground truth ↓ Change	Ground truth ↓ No-change	
Remote sensing → Change		TP	FP	TP + FP
Remote sensing → No-change		FN	TN	FN + TN
		TP + FN	FP + TN	TP+FN+FP+TN

Popular measures e.g.

Sensitivity = Producer's accuracy = $\frac{TP}{TP + FN}$

Prevalence = $\frac{TP + FN}{TP + FN + FP + TN}$

Others (e.g. user's accuracy) may be derived.

A simple question?

How accurate is this classification (or estimates of change)?

		Ground truth ↓ Change	No-change	
Remote sensing →	Change	160	160	320
	No-change	100	580	680
		260	740	1000

Is the producer's accuracy = 160/260 → 61% ?

Is amount of change = 260/1000 → 26% ?

NO, because the matrix might look like

		Ground truth ↓ (Se=Sp=0.90)		
		Change	No-change	
Remote sensing → (Se=Sp=0.80)	Change	160	160	320
	No-change	100	580	680
		260	740	1000

but is actually:

		Ground truth ↓ TRUTH		
		Change	No-change	
Remote sensing → (Se=Sp=0.80)	Change	160	160	320
	No-change	40	640	680
		200	800	1000

Occurs because ground data set is imperfect:

		Ground truth ↓ TRUTH: Change	Ground truth ↓ TRUTH: No-change	
Ground truth → (Se=Sp=0.90)	Change	180	80	260
	No-change	20	720	740
		200	800	1000

Good news - can correct for ground data error.

Note - here assumed conditional independence (trends more complex and can be in different direction if invalid and will be invalid in many studies).

Impact on estimation

Real accuracy (%)		Perceived	
Ground data	Remote sensing	RS accuracy	Prevalence
90	80	61	26
95	90	76	23

Systematically underestimate accuracy of remote sensing change detection and overestimate amount of change.

Impact of imperfect ground data

➔ **Systematic** bias.

e.g.

- Underestimate producer's accuracy.
- Typically overestimate prevalence (e.g. amount of change).

Magnitude of bias can be very large for even if ground data set is highly accurate.

Can correct/compensate for ground data error.

PART 2: Comparisons

Often compare (e.g. accuracy over time, change rates between regions). Based on comparison of proportions.

Must design an accuracy assessment programme to meet its objectives.

One key concern is the 

of the testing set.

Too large – any non-zero difference will appear statistically significant.

Too small – programme may fail to detect an important difference.

Sample size determination

Often based on precision to estimate proportion

$$p \pm h = p \pm z_{\alpha/2}(\mathrm{SE}) \qquad \mathrm{SE} = \sqrt{\frac{p(1-p)}{n}}$$

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

BUT

Aim is often not to estimate accuracy to a given precision but to use in a comparative analysis

- comparison against a target
- comparison against another accuracy

(e.g. classifier comparison)

….need to consider additional properties.

Comparison

Very common *v*. target e.g. 85%

or classifier evaluation e.g.

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

BUT

often inappropriate &

pessimistically biased →

International Journal of Remote Sensing
Vol. 29, No. 11, 10 June 2008, 3137–3158

Taylor & Francis
Taylor & Francis Group

Harshness in image classification accuracy assessment

GILES M. FOODY*
School of Geography, University of Nottingham, University Park, Nottingham NG7 2RD, UK

(*Received 1 September 2006; in final form 3 May 2007*)

Thematic mapping via a classification analysis is one of the most common applications of remote sensing. The accuracy of image classifications is, however, often viewed negatively. Here, it is suggested that the approach to the evaluation of image classification accuracy typically adopted in remote sensing may often be unfair, commonly being rather harsh and misleading. It is stressed that the widely used target accuracy of 85% can be inappropriate and that the approach to accuracy assessment adopted commonly in remote sensing is pessimistically biased. Moreover, the maps produced by other communities, which are often used unquestioningly, may have a low accuracy if evaluated from the standard perspective adopted in remote sensing. A greater awareness of the problems encountered in accuracy assessment may help ensure that perceptions of classification accuracy are realistic and reduce unfair criticism of thematic maps derived from remote sensing.

Comparative analysis

Comparative analyses often based on hypothesis testing.

e.g. H_o – no difference in accuracy
H_1 – the accuracy values differ

Two types of error:

Type I – when H_1 is incorrectly accepted (declare a difference as being significant when it is not).

Type II – when H_o is incorrectly accepted (fail to detect a meaningful difference that does exist).

Type I error

H_1 is incorrectly accepted (declare a difference as being significant when it is not).

Probability of making a Type I error is expressed as the significance level, α

Commonly set $\alpha = 0.05$
(i.e. a 5% chance of inferring a significant difference exists when actually is no difference)

Type II error

H_o is incorrectly accepted (fail to detect a meaningful difference that does exist).

Probability of making a Type II error is β and related to the power of the test $(1-\beta)$.

Type I errors typically viewed x4 more important than Type II, so commonly, $\beta = 0.2$

or $(1-\beta) = 0.8$

If $(1-\beta) = 0.8$ – 80% chance of declaring a difference that exists as being significant.

Is 0.8 adequate ?

Many studies often fail to detect a significant difference – did the study have sufficient power?

Tests using small sample sizes often underpowered.

Difficult to interpret non-significant results (is there really no difference or just failed to identify it?)

Estimating sample size

To determine sample size need to consider:

- Significance level α
- Power $(1-\beta)$
- Effect size – minimum meaningful difference.

e.g. common remote sensing scenario *v* target

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

and with continuity correction:

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

Use acquired data to test for difference using:

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

e.g. common scenario *v* another accuracy

$$n' = \frac{(z_{\alpha/2}\sqrt{2\overline{P}\,\overline{Q}} + z_{\beta}\sqrt{P_1Q_1 + P_2Q_2})^2}{(P_2 - P_1)^2}$$

and with continuity correction:

$$n = \frac{n'}{4}\left(1 + \sqrt{1 + \frac{4}{n'|p_2 - p_1|}}\right)^2$$

Use acquired data to test for difference using:

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

Note:

1. Equations may be re-written

e.g.

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

2. Can also use alternatives for related samples (e.g. McNemar test).

3. Instead of hypothesis testing could use confidence intervals.

So what?

Remember, important to use appropriate size

Too large – any non-zero difference will appear statistically significant.

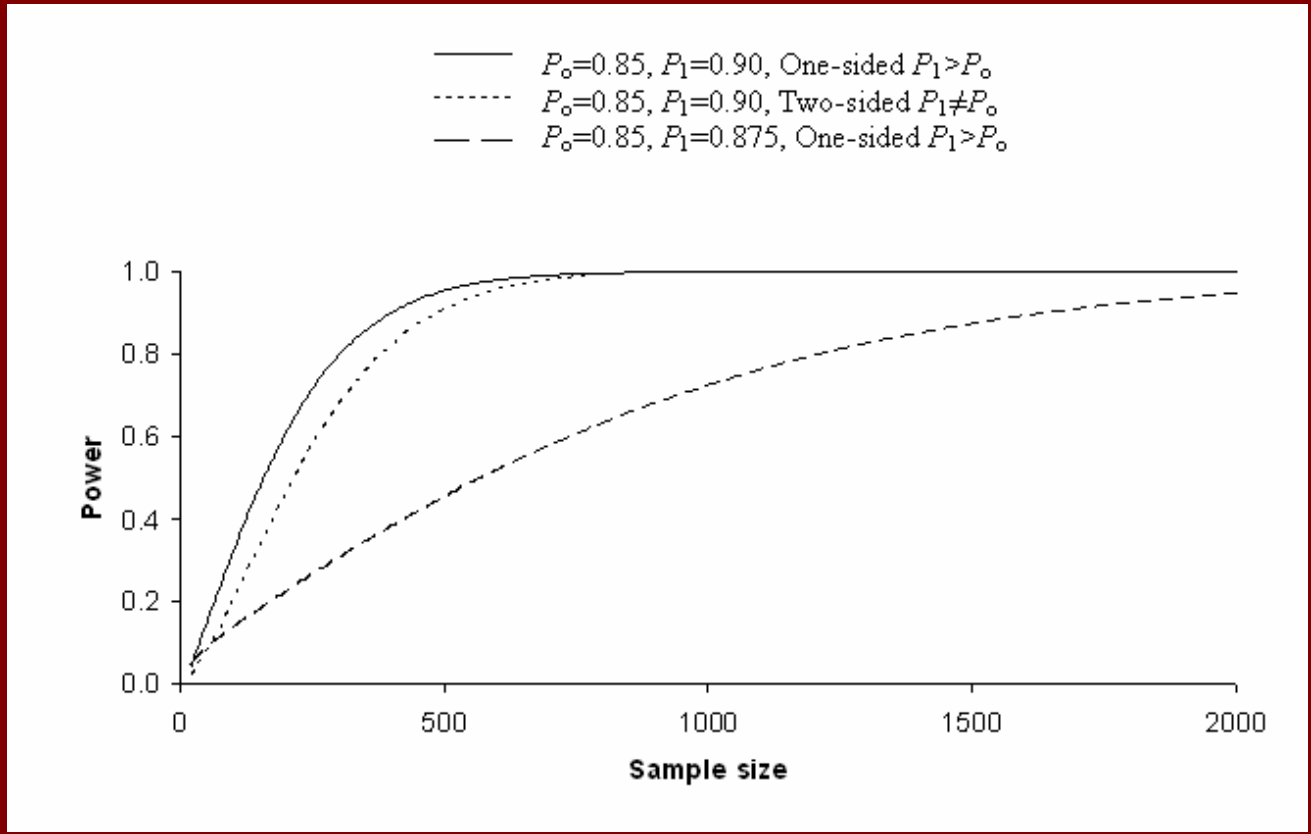
Too small – fail to detect an important difference.

Sizes used in remote sensing….

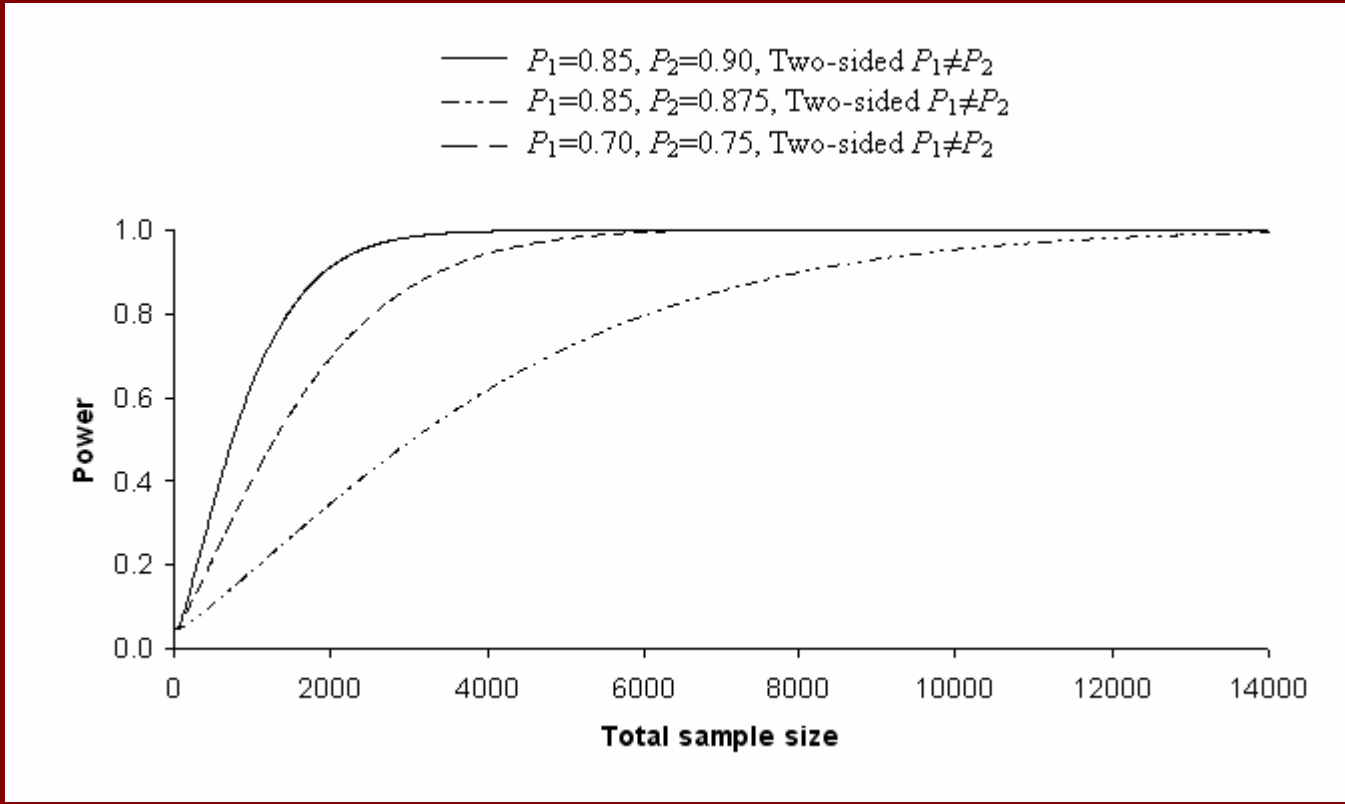
- range from 10s…100s…1000s…10,000+

Size needed:

ν target



v another accuracy



Conclusions

Error in ground 'truth' can lead to systematic bias – underestimates accuracy and is correctable.

Accuracy assessment often has a comparative component – has implications for sample size (need to specify effect size, α, and β).

Required size may be quite large. Need to be aware of danger of using inappropriate size (too small or too large).

☺ *The end*