Evaluating and Characterising Image Classification Interoperability

K. Lowell[1,2], A. Mitchell[3], I. Tapley[3], A. Milne[3], E. Lehmann[4], Z-S. Zhou[4], P. Cacetta[4], A. Held[5]

[1]CRC-SI, Department of Geomatics, University of Melbourne, Carlton, Victoria, Australia
[2]Dept. of Primary Industries-Victoria, Parkville, Victoria, Australia
[3]Cooperative Research Centre for Spatial Information (CRC-SI), School of Biological, Earth and Environmental Sciences, The University of New South Wales, Sydney, Australia
[4]CSIRO Mathematics, Informatics and Statistics, Floreat, Western Australia
[5]CSIRO AusCover Facility, Canberra, Australia

Abstract – Interoperability is generally thought of as being nominal ("interoperable/not interoperable") whereas in reality it should be assessed in an ordinal, interval, or ratio manner. This article presents a number of techniques that can be used to describe the interoperability between classifications in a way that is more robust than a confusion matrix. The Australian island state of Tasmania is used as the demonstration area. Interoperability of forest/non-forest classifications produced from radar (PALSAR) and a local map product were used to demonstrate evaluation techniques that can be used to characterise interoperability. It is concluded that these techniques allow a statistical and spatial evaluation of interoperability, and that the techniques can also be adapted to evaluating classifications. Finally, the techniques presented also provide information that can be used to modify classification procedures with a view to enhancing interoperability.

Keywords: TASVEG, PALSAR, radar, statistics, forest classification, Tasmania

1. INTRODUCTION

In recent years, a growing desire for image interoperability has developed. In addition to a general desire for consistency of output products and map interchangeability, this is also driven by the reality that the long-term continuity of most satellite sensors is not assured. However, it is well recognised that data from different sensors have application-specific strengths and weaknesses that make "perfect" or "true" or "total" interoperability a difficult goal.

The ultimate goal of perfect interoperability suggests that there is a single definition of, and objective for, interoperability that equates to two different sensors producing exactly the same results for all pixels.

With such an objective, interoperability is easy to assess – either all pixels are classified the same or they are not. Yet perfect interoperability is likely to be unachievable in all but the rarest cases. Moreover, pixel-based interoperability may not even be necessary depending on the use to which classified imagery is being put. If one adopts a more realistic concept and goal for interoperability, then more information about interoperability is necessary than simply "achieved/not achieved." Thus more flexibility in defining interoperability engenders greater complexity in assessing the interoperability between classifications produced by different sensors.

In addition to defining interoperability clearly, equally important is that the assessment of interoperability must be made relative to a particular goal. For example, Tier 1 reporting for international carbon accounting is based on a country-wide estimate of forest. Hence if two sensors produce the same forest estimate for an entire country, the two are interoperable for the purposes of Tier 1 reporting even if the internal locations of forest are completely different. This suggests the need for techniques for evaluating interoperability that provide a plethora of information to inform a variety of interoperability definitions and objectives.

Nonetheless, there is little need to conduct specialised evaluation for each interoperability definition and imagery use. Instead, it is possible to define a suite of analytical techniques that will produce information that is useful for a range of interoperability definitions and imagery uses. The goal of this paper is to present a number of individual measures that can contribute to this objective, and to discuss how their collective interpretation provides for non-binary evaluation of interoperability among maps that may have been produced using data from two different sensors.

2. DATA AND STUDY AREA

The focus area of this work is the Australian island state of Tasmania (Figures 1 and 3) that lies 240 kilometres south of the eastern part of the Australian mainland and that covers some 68,000 square kilometres. The north western and eastern portions of Tasmania are predominantly forested with agriculture being prominent in the central zone. Coastal scrub/buttongrass dominates the western margins.

For 2007, 46 ALOS PALSAR images captured from August to October were obtained from the Japanese space agency JAXA. A forest/non-forest Tasmania-wide mosaic was produced from the 2007 radar imagery (Figure 1). A schematic of the way that these radar data were processed to produce the 2007 mosaic is presented in Figure 2.

The map used to assess interoperability is known as TASVEG. TASVEG categorizes vegetation into 147 communities across the entirety of Tasmania. It is produced through human interpretation of (1:25,000 and 1:42,000) aerial photographs. It was first produced in 1998 and has been constantly updated ever since. The TASVEG map employed herein was acquired in early 2010 and the 147 vegetative communities placed into the forest or non-forest class (Figure 3).

The use of a map updated in 2010 for comparison with a 2007 PALSAR classification in an area that is known to have experienced some anthropogenic land cover change makes it a foregone conclusion that the two maps will not be perfectly interoperable. Moreover, of course, there is little interest in assessing the interoperability of classifications produced from digital image data with maps like TASVEG that have been developed through human interpretation of aerial photographs. However, the goal of this paper is to illustrate methodology that can be used for the evaluation of interoperability, rather than undertaking the actual assessment of the interoperability of two data sources; the use of TASVEG is adequate for this purpose. Moreover, actual work being undertaken to assess the interoperability of two digital image data sources is subject to confidentiality agreements and cannot be published herein.

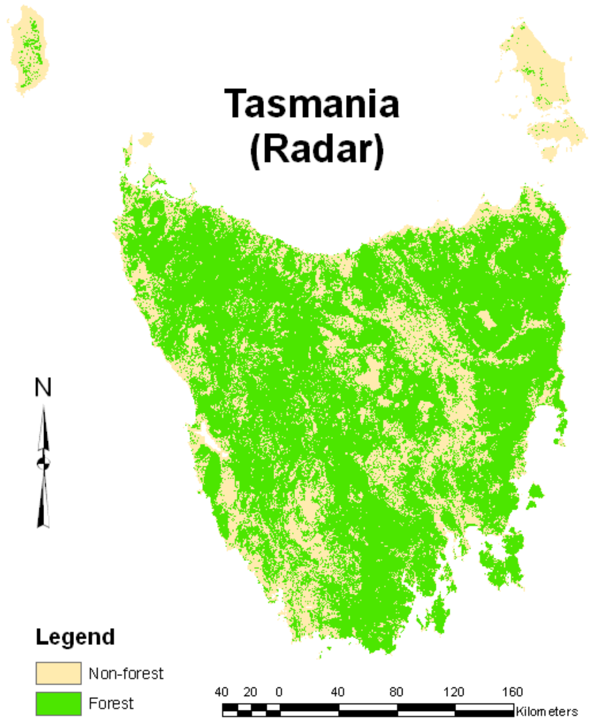
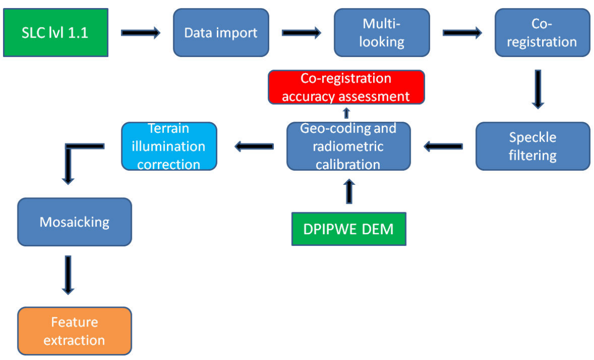


Figure 1. Tasmania classified into Forest/Non-forest using PALSAR data.

Figure 2. Radar processing sequence.



3. ASSESSING INTEROPERABILITY

Visual comparison of Figures 1 and 3 clearly indicates that the PALSAR and TASVEG classifications are not perfectly interoperable; again, for the purposes of methodology development, the source of the classifications, the magnitude of the difference, and the accuracy of each classification is of no importance. This can be further shown by creating a map that shows any areas of difference (Figure 4) with the results tabulated in a conventional confusion matrix (Table 1; Congalton and Green 1999). However, this gives a limited amount of information to characterise the interoperability. For example, there is no quantification of the spatial distribution, or statistical significance of differences[1].

Figure 3. Tasmania classified into Forest/Non-forest using TASVEG.

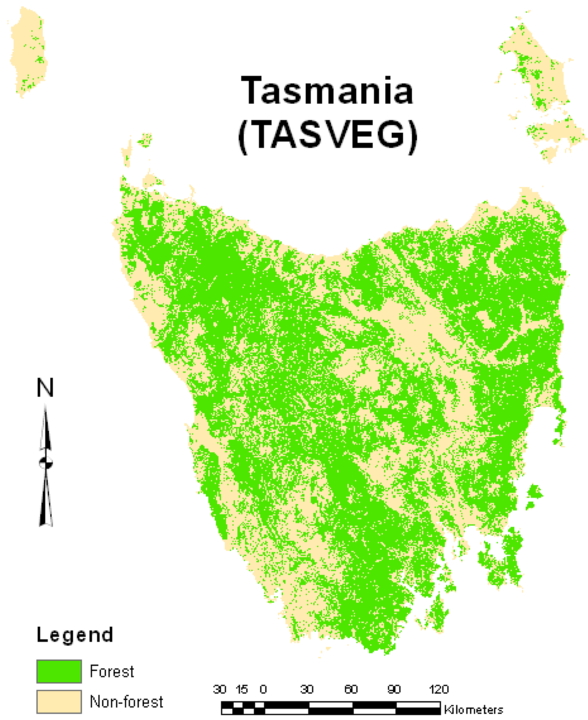


Table 1. Confusion matrix for radar vs. TASVEG (Values are in ha).

		TASVEG		
		Non-forest	**Forest**	**Total**
Radar	**Non-forest**	2,203,512 (32%)	287,825 (4%)	2,491,337 (36%)
	Forest	703,056 (10%)	3,651,116 (53%)	4,354,173 (64%)
	Total	2,906,568 (42%)	3,938,941 (58%)	

To provide more interoperability information than is available in a confusion matrix, obtaining a sense of the variability in differences is critical. To achieve this, 55 10-km-by-10-km (10,000 ha) samples were established across Tasmania on a square grid with samples spaced 33 km apart; this approach was developed by Lowell (2001) and adapted for operational use by Barson *et al*. (2004).

The first means of assessment is a statistical evaluation of the mean difference between the samples (Table 2); because a binary forest/non-forest map is being evaluated, the analysis is only shown for one class. This statistical evaluation provides a clear indication that globally the two data sources

[1] It is noted that the confusion matrix can be tested for significance from randomness through the calculation of the kappa coefficient and its standard error.

have produced different maps. This alone suggests a complete lack of global interoperability. Hence even if one were only interested in having the same estimate for the amount of forest and non-forest across Tasmania – the coarsest level of interoperability – TASVEG and radar cannot be considered interoperable.

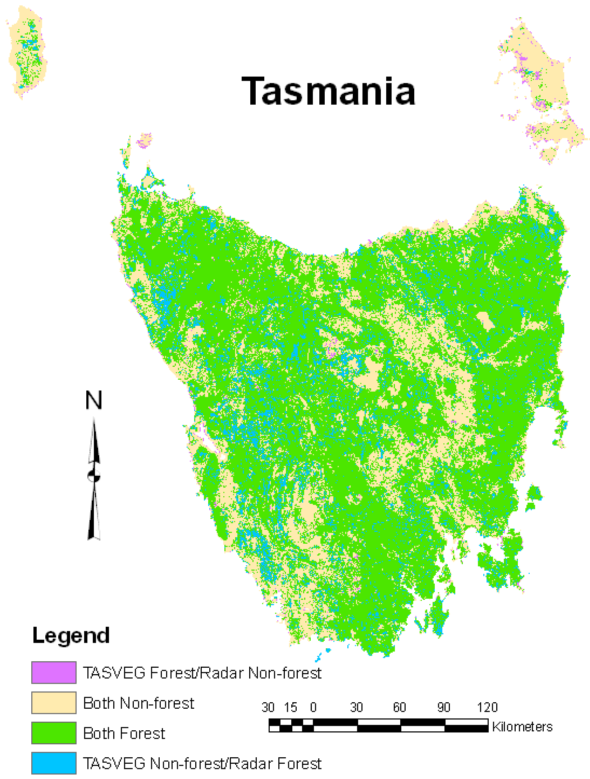


Figure 4. Differences between TASVEG and radar forest/non-forest classifications for Tasmania.

Table 2. Global statistical analysis of the amount of Non-forest (in ha) from 55 10,000 ha samples from TASVEG and radar classifications.

	TASVEG	Radar	Difference
Minimum	573	234	-1346
Maximum	9830	9922	2231
Mean	4214	3642	572
Std. Dev.	2401	2342	706
Std Err.	324	316	95
		Student's *t*	6.014
		p	0.000

It is doubtful, however, that even if the analysis presented in Table 2 had shown no statistically significant difference, that the classifications from the two data sources are not perfectly interoperable at the pixel level.

One way of assessing this is through an examination of the frequency distribution of the forest/non-forest for each sample unit (Figure 5). This shows clearly (the left of Figure 5) that the radar classification has many more samples on which a smaller area of non-forest is estimated to be present than does TASVEG. However, the differences across the frequency distribution do not indicate the relationship of the non-forest for individual sample units.

This can be assessed using regression analysis (Figure 6; Table 3). Table 3 indicates that the correlation between the amount of non-forest on the TASVEG and radar maps is statistically significant, and that the regression line is not significantly different from the ideal line where the intercept is 0.0 and the slope is 1.0. However, the root-mean-square-error (RMSE) is 693. ha or about 7% of the sample unit size of 10,000 ha. Figure 6 also clearly demonstrates the bias demonstrated by Table 2 with the regression line consistently below the ideal line.

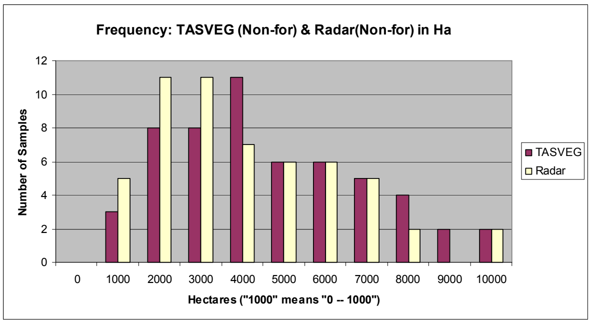


Figure 5. Frequency distribution for Non-forest for 55 areal sample units.

This regression analysis also provides additional information about interoperability. Certain samples are poorly estimated – those farthest from the red line – whereas others are estimated reasonably well. If the latter are spatially contiguous, there may be regions where a high level of interoperability is achieved.

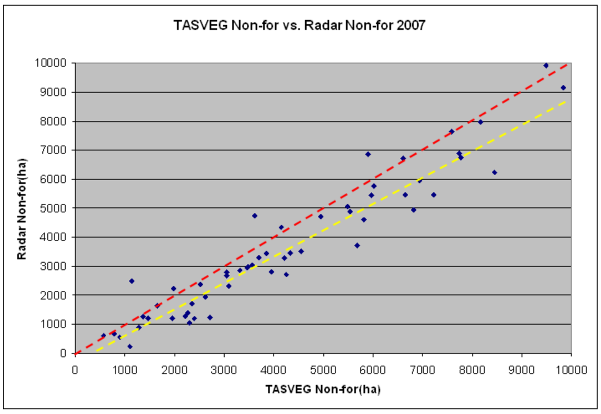


Figure 6. Comparison of Non-forest estimated by TASVEG and radar on 10,000 ha areal samples. Red line is the ideal regression line (Intercept = 0.0 and Slope = 1.0); yellow line is actual regression line.

Table 3. Regression analysis associated with Figure 6. Values in parentheses are *p* values for overall significance of correlation (R^2), difference from 0.0 (Intercept), and difference from 1.0 (Slope).

	Adj. R^2	Intercept	Slope	RMSE
All obs. (n=55)	0.912 (0.000)	-287. (0.137)	0.932 (0.091)	693. ha

Figure 7 shows the spatial distribution of the differences between the TASVEG and radar non-forest classifications. The orange class (-75 to +75) is the one for which the differences are the smallest. That this class does not appear to cluster in any noticeable pattern suggests that there is no spatially consistent pattern to small differences, and therefore no particular region where localised interoperability might be higher. Conversely, the northeast of Tasmania shows differences that are relatively high indicating that in this region there is a lack of localised interoperability with a tendency for TASVEG to identify a greater amount of non-forest than radar.

In identifying spatial distributions of differences, localised assessment of interoperability can be undertaken through visual examination of ancillary data. Figure 8 shows the TM imagery, and TASVEG and radar classifications for the sample unit for which the difference was the smallest (52 ha or 0.5% of the 10,000 ha sample size). For this sample pixel-by-pixel interoperability has almost been achieved. While this may seem obvious with such a small difference, in fact having a small regional difference does not guarantee localised reliability.

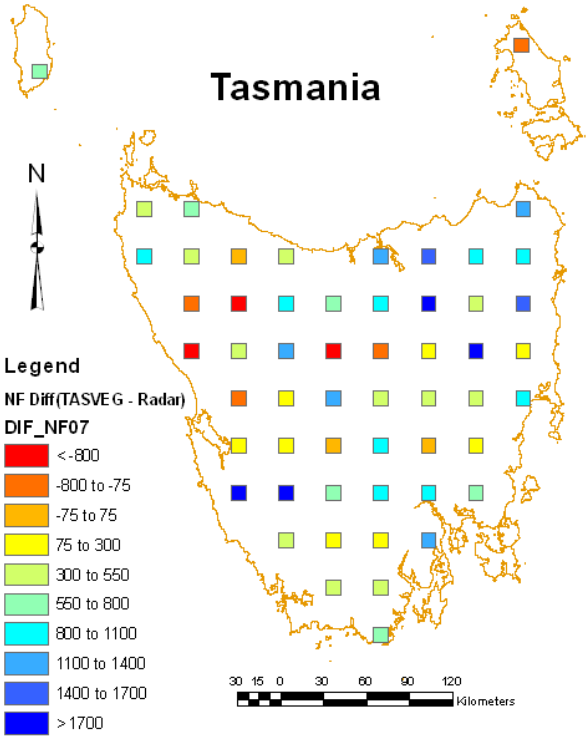


Figure 7. Spatial distribution of differences between TASVEG and radar on the 55 areal samples extracted.

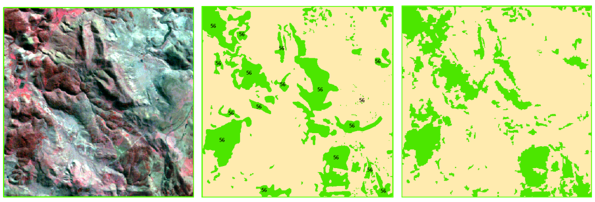


Figure 8. TM image and TASVEG and radar forest/non-forest classification for areal sample with smallest difference.

4. DISCUSSION AND CONCLUSIONS: IMPLICATIONS FOR INTEROPERABILITY

This paper has demonstrated a number of ways to evaluate interoperability based on an evaluation of classified image products. The demonstration was facilitated by using two data sources that showed obvious differences in classifications. The example presented was also simplified because it employed a binary taxonomy. Nonetheless, more complex taxonomies can be accommodated by the methods presented. However, regardless of the techniques employed an increased number of classes makes it difficult to assess interoperability for all classes. Similarly, interoperability of relatively small classes, regardless of the total number of classes in the taxonomy and methodology employed is difficult to assess.

The techniques presented were used to assess interoperability rather than classification accuracy. In assessing interoperability, the focus is on differences between classifications and not comparison against a particular standard. Evaluating accuracy requires that one classification or data source be acknowledged as the one of higher accuracy and/or requires the availability of ancillary data that are accepted as being authoritative. If such data are available, the techniques presented can be used to evaluate classification accuracy while obtaining more detailed information than what is produced by a conventional confusion matrix.

One of the benefits of the techniques employed is that they provide a road map for increased interoperability. The information presented provides an understanding of the magnitude and characteristics of the interoperability. Regression analysis (Table 3; Figure 6) indicates the level of difference that can be expected and the spatial representation (Figure 7) provides a means of assessing if there are regions where interoperability is particularly high or low. The northeast of Tasmania was identified as an area where differences between TASVEG and radar classifications were particularly high. Ancillary knowledge of topography, soils, or other factors in this region may allow targeting of classification modifications that would increase interoperability.

REFERENCES

M. Barson, V. Bordas, K. Lowell, K. and K. Malafant, “Independent reliability assessment for the Australian Agricultural Land Cover Change Project 1990/91-1995, Remote Sensing and GIS Accuracy Assessment,” CRC Press, 2004.

R. Congalton, and K. Green, “Assessing the Accuracy of Remotely Sensed Data: Principles and Practices,” Lewis Publishers, 1999.

Lowell, K., “An area-based accuracy assessment methodology for digital change maps,” International Journal of Remote Sensing, vol. 22(17), p.p. 3571-3596, 2001.