

## ANEXO 5. ARTÍCULOS DE REFERENCIA

---

**ANEXO 5.1.** NORMALIZACIÓN RADIOMÉTRICA CON IR-MAD, ITERATIVELY REWEIGHTED MULTIVARIANTE ALTERATION DETECTION (CANTY Y NIELSEN, 2008)

**ANEXO 5.2.** RELLENO DE VACÍOS CAUSADOS POR MÁSCARA DE NUBES Y SOMBRAS (MARTINUZZI ET AL. 2007)

**ANEXO 5.3.** GOOD PRACTICES FOR ESTIMATING AREA AND ASSESING ACCURACY OF LAND CHANGE (OLOFSSON ET AL. 2014)

ANEXO 5.1. NORMALIZACIÓN RADIOMÉTRICA CON IR-MAD,  
ITERATIVELY REWEIGHTED MULTIVARIANTE ALTERATION  
DETECTION (CANTY Y NIELSEN, 2008)

---

# Automatic radiometric normalization of multitemporal satellite imagery with the iteratively re-weighted MAD transformation

Morton J. Canty<sup>a,\*</sup>, Allan A. Nielsen<sup>b,1</sup>

<sup>a</sup> *Institute for Chemistry and Dynamics of the Geosphere, Jülich Research Center, D-52425 Jülich, Germany*

<sup>b</sup> *Danish National Space Center, Technical University of Denmark, Richard Petersens Plads, Building 321, DK-2800 Kgs. Lyngby, Denmark*

Received 27 March 2007; received in revised form 21 June 2007; accepted 15 July 2007

## Abstract

A recently proposed method for automatic radiometric normalization of multi- and hyperspectral imagery based on the invariance property of the Multivariate Alteration Detection (MAD) transformation and orthogonal linear regression is extended by using an iterative re-weighting scheme involving no-change probabilities. The procedure is first investigated with partly artificial data and then applied to multitemporal, multispectral satellite imagery. Substantial improvement over the previous method is obtained for scenes which exhibit a high proportion of change.

© 2007 Elsevier Inc. All rights reserved.

*Keywords:* Change detection; Multispectral imagery; Iteratively reweighted multivariate alteration detection (IR-MAD); Canonical correlation analysis (CCA); Radiometric

## 1. Introduction

Ground reflectance determination from satellite imagery requires, among other things, an atmospheric correction algorithm and the associated atmospheric properties at the time of image acquisition. For most historical satellite scenes such data are not available and even for planned acquisitions they may be difficult to obtain. A relative normalization using the radiometric information intrinsic to the images themselves is an alternative whenever absolute surface reflectances are not required.

In performing relative radiometric normalization, one usually makes the assumption that the relationship between the at-sensor radiances recorded at two different times from regions of constant reflectance can be approximated by linear functions. The critical aspect is the determination of suitable time-invariant features upon which to base the normalization (Du et al., 2002; Furby and Campbell, 2001; Hall et al., 1991; Moran et al., 1992; Schott et al., 1988; Yang and Lo, 2000).

In Canty et al. (2004) a fully automatic procedure was suggested for determining time-invariant observations which takes advantage of the invariance properties of the Multivariate Alteration Detection (MAD) transformation (Nielsen et al., 1998). The method was validated with unbiased statistical tests for equal means and variances of the normalized image bands and also compared quantitatively to manual selection of pseudo invariant features for normalization.

Since then, an iteratively re-weighted modification of the MAD transformation (IR-MAD) has been introduced (Nielsen, 2007) which establishes a better background of no change upon which to examine significant changes. The IR-MAD procedure is superior to the ordinary MAD transformation in identifying significant change, particularly for data sets in which the fraction of invariant pixels is relatively small, e.g., scenes which undergo large seasonal changes in vegetation or land use. It should therefore also perform better in isolating the no-change pixels suitable for use in relative radiometric normalization. In the present investigation this has been confirmed by comparing radiometric normalizations obtained with the original and with the iterated versions of the MAD transformation. Since the original MAD procedure has already been demonstrated to lead to as good or better normalizations than those achieved with manually determined invariant features (Canty et al., 2004; Schmidt et al.,

\* Corresponding author. Tel.: +49 2461 614885; fax: +49 2461 612496.

E-mail addresses: [m.canty@fz-juelich.de](mailto:m.canty@fz-juelich.de) (M.J. Canty), [aa@space.dtu.dk](mailto:aa@space.dtu.dk) (A.A. Nielsen).

<sup>1</sup> Tel.: +45 4525 3425; fax: +45 4588 1397.

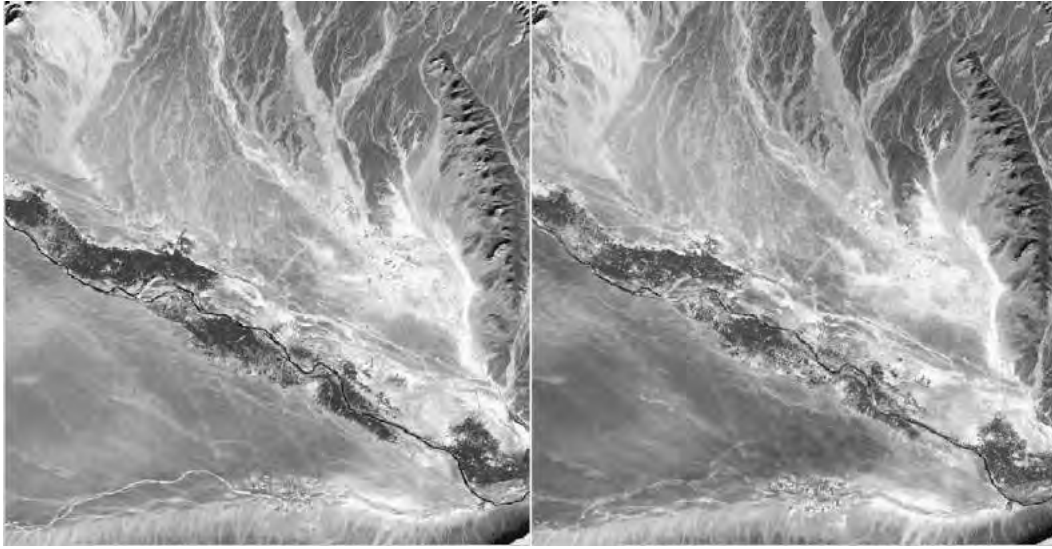


Fig. 1. Landsat ETM+ images over east central Morocco, left: December 19, 1999, right: October 18, 2000.

2005; Schroeder et al., 2006), further comparison with such methods is unnecessary.

Singh (1989) gives a good but now somewhat outdated survey of change detection algorithms for remotely sensed data. For recent reviews in a more general context see Radke et al. (2005) or Coppin et al. (2004). In Section 2 below we focus on the IR-MAD change detection procedure and briefly summarize its properties. The satellite imagery used for the comparison is described in Section 3. In Section 4 we apply the IR-MAD transformation to find invariant pixels for a partly artificial data set. In Section 5 results are described for a bitemporal arid scene in which relatively little change occurs, as well as for several image pairs exhibiting very substantial changes in ground reflectance. Invariant pixels obtained from IR-MAD are used to perform the necessary regressions for relative radiometric normalization and the quality of the normalized images is evaluated in terms of paired *t*-tests and *F*-tests for equal means

and variances, respectively. Results are discussed and conclusions are drawn in Section 6.

## 2. The MAD and IR-MAD transformations

Consider first of all two *N*-band multispectral images of the same scene acquired at different times, between which acquisitions ground reflectance changes have occurred at some locations, but not everywhere. Assume without loss of generality both images to have pixel intensities with zero mean. Representing observations (pixel intensities) in the first multispectral image by a random vector  $F = (F_1, \dots, F_N)^T$ , we can make a linear combination of the intensities for all spectral bands, creating a scalar image characterized by the random variable  $U = \mathbf{a}^T F$ . The vector of coefficients  $\mathbf{a}$  is as yet unspecified. We do the same for the second image, represented by  $G$ , forming the linear combination  $V = \mathbf{b}^T G$ , and then look at the scalar difference image  $U - V$ . This combines

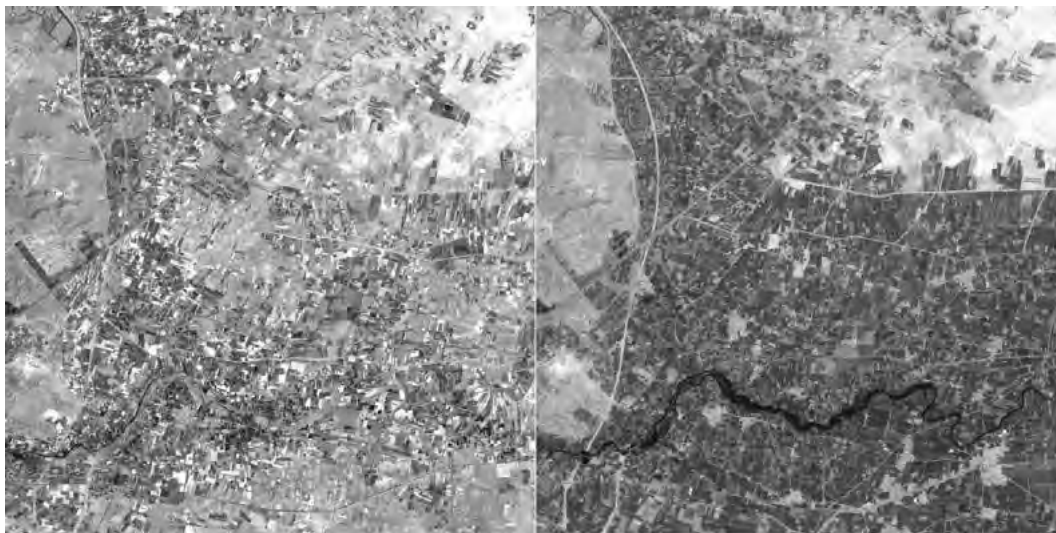


Fig. 2. ASTER images near Esfahan, Iran, left: July 30, 2001, right: May 22, 2005.

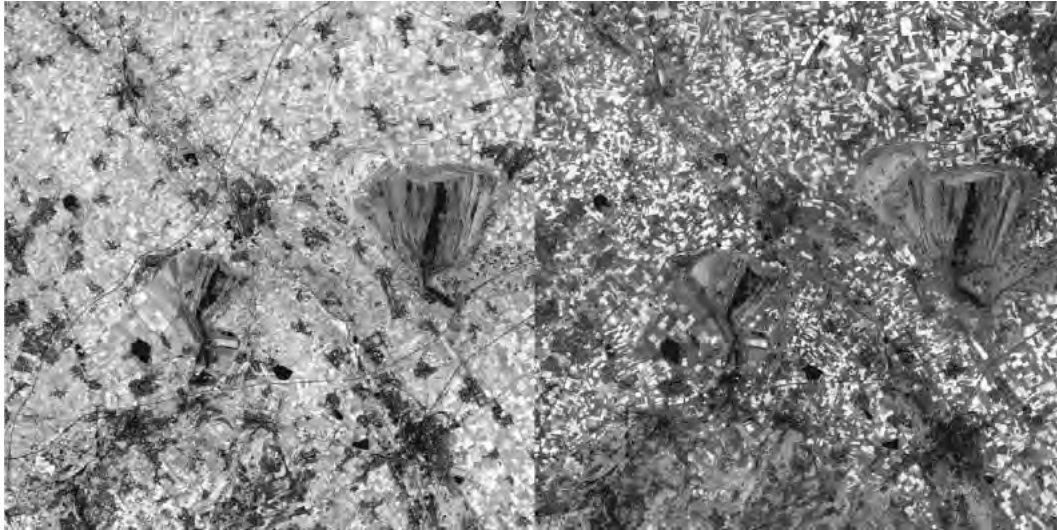


Fig. 3. Landsat ETM+ images over Jülich, Germany, left: June 26, 2001, right: August 29, 2001.

all of the change information into a single image, but one has of course still to choose the coefficients  $\mathbf{a}$  and  $\mathbf{b}$  in some suitable way. In Nielsen et al. (1998) it is suggested that they be determined by applying standard Canonical Correlation Analysis (CCA), first described by Hotelling (1936), to the images  $F$  and  $G$ . This leads to the coupled generalized eigenvalue problems

$$\begin{aligned} \Sigma_{fg} \Sigma_{gg}^{-1} \Sigma_{gf} \mathbf{a} &= \rho^2 \Sigma_{ff} \mathbf{a} \\ \Sigma_{gf} \Sigma_{ff}^{-1} \Sigma_{fg} \mathbf{b} &= \rho^2 \Sigma_{gg} \mathbf{b}, \end{aligned} \quad (1)$$

where  $\Sigma_{ff}$  and  $\Sigma_{gg}$  are the covariance matrices of the two images and  $\Sigma_{fg} = \Sigma_{gf}^T$  is the inter-image covariance matrix. Solution of the eigenvalue problems (Eq. (1)) generates new multispectral images  $\mathbf{U} = (U_1, \dots, U_N)^T$  and  $\mathbf{V} = (V_1, \dots, V_N)^T$ , the components of which are called the canonical variates (CVs). The CVs are ordered by similarity (correlation) rather than, as in the original images, by wavelength. The canonical correlations  $\rho_i = \text{corr}(U_i, V_i)$ ,  $i = 1, \dots, N$ , are the square roots of the eigenvalues of the coupled eigenvalue problem and  $\mathbf{a}_i$  and  $\mathbf{b}_i$ ,  $i = 1, \dots, N$ , which determine  $\mathbf{U}$  and  $\mathbf{V}$  from  $F$  and  $G$ , are the eigenvectors. The pair  $U_1, V_1$  is maximally correlated, the pair

$(U_2, V_2)$  is maximally correlated subject to being orthogonal to (uncorrelated with) both  $U_1$  and  $V_1$ , and so on. Performing paired differences (in reverse order) then generates a sequence of transformed difference images

$$M_i = U_{N-i+1} - V_{N-i+1}, \quad i = 1 \dots N, \quad (2)$$

referred to as the MAD variates, having nice statistical properties which make them very useful for visualizing and analyzing change information. Thus for instance they are uncorrelated,

$$\text{cov}(M_i, M_j) = 0, \quad i \neq j, \quad (3)$$

and have, by virtue of our chosen ordering, successively decreasing variances given by

$$\text{var}(M_i) = \sigma_{M_i}^2 = 2(1 - \rho_{N-i+1}^2). \quad (4)$$

The fact that several additions and subtractions go into determining the MAD variates means that, from the Central Limit Theorem, they are ideally normally distributed. The usefulness of the MAD variates for automatic radiometric

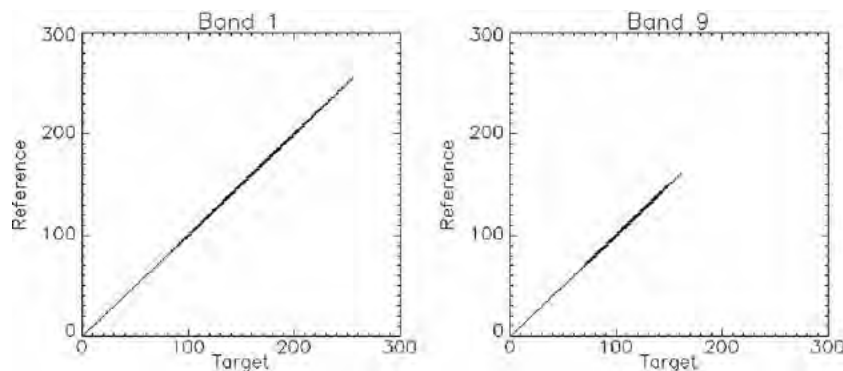


Fig. 4. IR-MAD normalization: Orthogonal regressions on spectral bands 1 and 9 for a 10% simulated no-change fraction in the ASTER bitemporal image. Only pixels identified as unchanged are plotted.

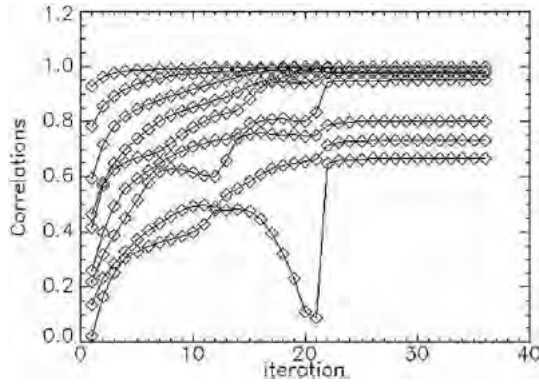


Fig. 5. Canonical correlations for the ASTER images with simulated no-change pixels.

normalization stems from the fact that they are invariant under affine transformations of either or both of the original images. This invariance is demonstrated explicitly in Canty et al. (2004).

Now consider two images of a scene, acquired at different times under similar illumination conditions, but for which no ground reflectance changes have occurred whatsoever. Then the only differences between them will be due to random effects like instrument noise and atmospheric fluctuation. In such a case we would expect that the histogram of any difference component that we generate would be very nearly Gaussian. In particular, the MAD variates, being uncorrelated, should follow a multivariate normal distribution with diagonal covariance matrix.

The MAD variates associated with change observations will deviate more or less strongly from such a multivariate normal distribution. Therefore, in the presence of genuine change, we expect an improvement of the sensitivity of the MAD transformation if we place emphasis on establishing an increasingly better background of no change against which to detect change. This can be done in an iteration scheme in which observations are weighted by the probability of no change, as determined on the preceding iteration, when estimating the sample means and covariance matrices which determine – via CCA – the MAD variates for the next iteration.<sup>2</sup>

The probability weights may be obtained by examining the MAD variates directly. Let the random variable  $Z$  represent the sum of the squares of the standardized MAD variates:

$$Z = \sum_{i=1}^N \left( \frac{M_i}{\sigma_{M_i}} \right)^2, \quad (5)$$

where  $\sigma_{M_i}$  is given by Eq. (4). Then, since the no-change observations are normally distributed and uncorrelated, their

realizations should be chi-square distributed with  $N$  degrees of freedom (distribution function  $P_{\chi^2;N}(z)$ ). For each iteration, the observations  $z$  (realizations of the random variable  $Z$ ) can then be given weights determined by the chi-square distribution, namely

$$\Pr(\text{no change}) = 1 - P_{\chi^2;N}(z). \quad (6)$$

$\Pr(\text{no change})$  is the probability that a sample  $z$  drawn from the chi-square distribution could be that large or larger. A small  $z$  implies a large probability of no change. Other weighting schemes are possible, for instance using unsupervised classification of change/no-change observations (Canty and Nielsen, 2006).

Iteration of the MAD transformation will continue until some stopping criterion is met, such as lack of change in the canonical correlations  $\rho_i$ ,  $i=1, \dots, N$ . In his original formulation of the iterative procedure, Nielsen (2007) standardized the MAD variates using the standard deviations of all of the observations, rather than using the iteratively re-weighted  $\sigma_{M_i}$  values given by Eq. (4). We prefer here to use re-weighted values, as this was found to lead to considerably better performance for normalizing bitemporal scenes in which change dominates.

For radiometric normalization purposes, one can select all pixels which satisfy  $\Pr(\text{no change}) > t$  where  $t$  is a decision threshold, typically 95%. The preferred method for performing the actual normalizations is then to determine the necessary transformations from orthogonal – as opposed to ordinary – linear regression on the invariant pixels, as demonstrated by Canty et al. (2004). A fraction of the invariant pixels may be withheld from the regression fit and used to test the normalized and reference images for equal radiometric characteristics.

In summary, the steps involved in radiometric normalization are as follows:

- Set weights equal to one for all pixels in the bitemporal scene.
- Repeat until canonical correlations cease to change significantly:
  - Do a weighted sample the bitemporal image to determine its mean vector and covariance matrix.
  - Perform CCA and construct the MAD variates  $M_i$ ,  $i=1, \dots, N$ .
  - Recalculate the weights according to Eqs. (5) and (6).

Table 1

Comparison of means and variances for 7850 hold-out test pixels, with paired  $t$ -tests and  $F$ -tests for equal means and variances of the Morocco images, normalization using the original MAD transformation

	Band 1	Band 2	Band 3	Band 4	Band 5	Band 7
Target mean	62.684	61.460	83.761	64.462	88.000	80.009
Ref. mean	75.509	68.501	91.131	73.193	98.775	90.335
Norm. mean	75.518	68.508	91.143	73.198	98.776	90.327
$t$ -stat.	-0.144	-0.0688	-0.0672	-0.036	-0.007	0.054
$P$ -value	0.885	0.945	0.945	0.975	1.000	0.957
Target var.	10.61	28.83	87.78	55.46	95.24	59.63
Ref. var.	17.10	42.72	129.64	90.83	129.82	81.98
Norm. var.	17.54	43.34	131.16	91.70	131.80	83.42
$F$ -stat.	1.025	1.014	1.011	1.009	1.015	1.017
$P$ -value	0.255	0.524	0.606	0.672	0.504	0.440

<sup>2</sup> There is a close analogy here to the application of iterated principal component analysis (PCA) to bitemporal images for univariate change detection, see, e.g. Wiemker (1997). After one PCA transformation the first principal axis only corresponds approximately to the (highly correlated) no-change pixels, since it is determined by both no-change and change observations. Weighting the observations inversely to their distance from the first principal axis and repeating the transformation improves the position of the principal axes relative to the no-change pixels and, correspondingly, enhances the change signal.

Table 2

As Table 1, but with 6671 hold-out test pixels and normalization after iteration of the MAD transformation to convergence

	Band 1	Band 2	Band 3	Band 4	Band 5	Band 7
Target mean	62.617	61.269	83.632	64.305	87.021	79.314
Ref. mean	75.535	68.498	91.339	73.186	98.097	89.940
Norm. mean	75.543	68.479	91.321	73.174	98.073	89.915
<i>t</i> -stat.	-0.554	1.055	0.686	0.604	1.063	1.169
<i>P</i> -value	0.580	0.292	0.493	0.546	0.288	0.243
Target var.	9.98	26.36	79.34	50.22	93.64	58.14
Ref. var.	14.94	37.70	116.46	79.96	115.39	70.65
Norm. var.	15.25	38.11	116.71	80.38	115.09	70.31
<i>F</i> -stat.	1.021	1.011	1.002	1.005	1.003	1.005
<i>P</i> -value	0.392	0.654	0.931	0.828	0.916	0.844

Table 3

Comparison of means and variances for 7700 hold-out test pixels, with paired *t*-tests and *F*-tests for equal means and variances of the Iran images, normalization using the original MAD transformation

	Band 1	Band 2	Band 3N	Band 4	Band 5	Band 6	Band 7	Band 8	Band 9
Target mean	83.105	95.208	129.564	119.392	111.495	118.465	109.393	105.607	101.404
Ref. mean	190.244	177.438	127.689	129.593	129.235	136.017	123.814	124.005	130.920
Norm. mean	190.206	177.382	127.707	129.591	129.230	136.001	123.805	124.005	130.899
<i>t</i> -stat.	0.240	0.340	-0.348	0.029	0.094	0.289	0.162	0.000	0.391
<i>P</i> -value	0.810	0.734	0.728	0.977	0.926	0.772	0.871	1.000	0.696
Target var.	179.39	349.25	153.34	326.10	333.29	407.37	339.39	365.58	268.70
Ref. var.	587.00	607.97	166.91	98.32	90.87	113.23	97.09	146.29	96.90
Norm. var.	704.39	680.72	168.75	82.62	80.89	99.69	83.88	123.50	85.86
<i>F</i> -stat.	1.120	1.120	1.011	1.190	1.123	1.136	1.157	1.185	1.129
<i>P</i> -value	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

- Select pixels with no-change probability (Eq. (6)) exceeding a threshold *t*.
- Perform an orthogonal regression on the selected pixels to determine the relative radiometric normalization coefficients (slope and intercept).

### 3. Imagery and software

Three sets of satellite images were used for the study: a bitemporal LANDSAT ETM+ scene (729 × 754 pixels) over an arid area in Morocco (Fig. 1, the same data as were used to demonstrate the original normalization method (Canty et al., 2004)), multi-temporal 1000 × 1000-pixel ASTER scenes over

an area southwest of Esfahan in Iran spanning a period of about four years (the set consisted of 5 image acquisitions in all, processed to “level 1b—registered radiance at the sensor”, two of which are shown in Fig. 2) and three LANDSAT ETM+ images (1000 × 1000 pixels) over the town of Jülich, Germany (Fig. 3 shows two of them). Extensive ground reference data were available for the Jülich acquisitions.

The image sets were registered to one another by applying an automatic contour matching algorithm (Li et al., 1995) and using first-order polynomial, nearest-neighbor resampling. The RMS errors were less than 0.5 pixel. In addition, the six short wave infrared (SWIR) bands of the ASTER images were sharpened to the 15m resolution of the three visual near infrared (VNIR) bands

Table 4

As Table 3, but with 8400 hold-out test pixels and normalization after iteration of the MAD transformation to convergence

	Band 1	Band 2	Band 3N	Band 4	Band 5	Band 6	Band 7	Band 8	Band 9
Target mean	84.569	97.784	125.644	124.262	118.328	126.174	114.979	109.349	107.488
Ref. mean	188.163	175.520	124.158	130.804	132.247	139.710	125.520	123.571	134.177
Norm. mean	188.204	175.511	124.155	130.802	132.243	139.701	125.515	123.553	134.161
<i>t</i> -stat.	-0.929	0.215	0.101	0.127	0.153	0.388	0.196	0.698	0.692
<i>P</i> -value	0.353	0.831	0.922	0.901	0.878	0.697	0.845	0.485	0.489
Target var.	50.067	80.689	101.680	52.979	30.015	40.225	30.887	68.753	19.037
Ref. var.	343.361	336.727	123.568	47.577	25.146	35.598	30.178	83.820	23.829
Norm. var.	352.881	342.440	122.949	46.742	24.620	35.138	29.591	83.054	24.157
<i>F</i> -stat.	1.028	1.017	1.005	1.018	1.021	1.013	1.020	1.009	1.01373
<i>P</i> -value	0.210	0.440	0.818	0.416	0.332	0.551	0.367	0.673	0.532

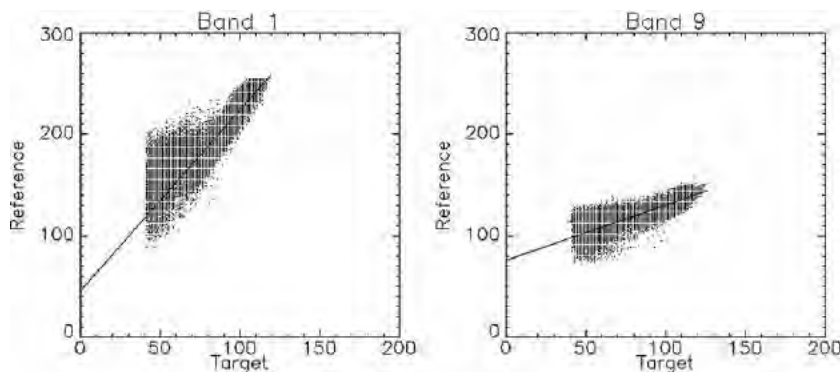


Fig. 6. Regressions on spectral bands 1 and 9 of the ASTER images for one iteration of the (i.e., the original) MAD transformation. Only pixels identified as unchanged are plotted.

with a wavelet fusion technique (Aiazzi et al., 2002). The ASTER images thus consisted of nine spectral bands each.

All image processing was performed within the ENVI remote sensing image analysis environment (ITT Visual Information Solutions). Extensions to ENVI for image registration, wavelet fusion, the IR-MAD transformation and radiometric normalization were written in the IDL language, see Canty (2007) for software availability.

#### 4. Simulated no-change observations

To illustrate the effectiveness of IR-MAD in correctly identifying no-change observations, an artificial data set was generated from two of the ASTER images (those shown in Fig. 2) by copying a  $316 \times 316 \times 9$  spatial/spectral subset from the upper left hand corner of the July 2001 image to the upper left hand corner of the May 2005 image, after adding a small amount of Gaussian noise (with standard deviation about 1% of the mean pixel intensity). The Gaussian noise prevents the solution of the coupled eigenvalue problem from becoming degenerate once the no-change pixels have been selected. The subset corresponds to a no-change fraction of 10% of the entire image.

The IR-MAD procedure identifies the copied pixels as no change to the exclusion of the rest of the scene. This is shown in Fig. 4, which displays the regression lines for radiometric

normalization for spectral bands 1 and 9. In the figure, *reference* refers to the image used as reference for the normalization and *target* is the image to be normalized. The regression lines go through the origin with  $45^\circ$  slope and are determined entirely by the duplicated pixels. Fig. 5 shows the convergence of the canonical correlations for the artificial data set. The large changes in the correlations up to about the 20th iteration correspond to a “zeroing in” on the artificial no-change observations. Thereafter convergence is fast. When the size of the copied subset was reduced below 10%, it was found that the iteration no longer converged to the duplicated pixels but rather to the physical no-change pixels. This is discussed in Section 6.

#### 5. Real data

We next investigated the effect of iterative re-weighting on radiometric normalization of a bitemporal arid scene, exemplified by the Morocco images, and of image pairs over Iran and Jülich which exhibited very significant seasonal and agricultural land-use changes.

##### 5.1. Morocco

The statistics for the normalization of the October 2000 LANDSAT ETM+ image over Morocco (target) to the December

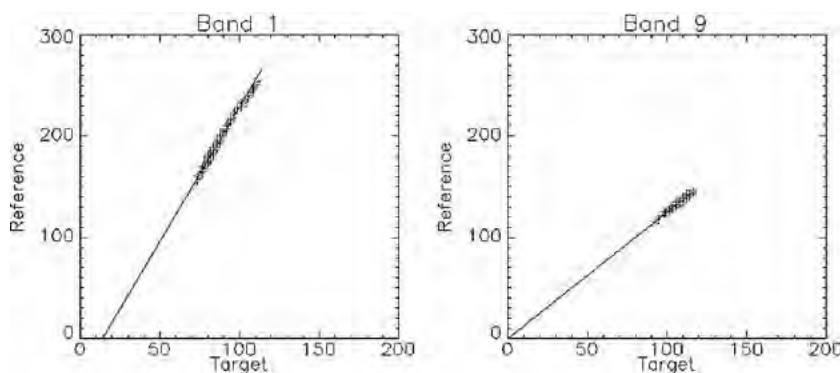


Fig. 7. Regressions on spectral bands 1 and 9 of the ASTER images after iteration of the MAD transformation to convergence. Only pixels identified as unchanged are plotted.



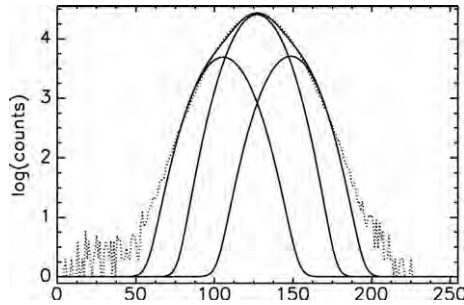


Fig. 8. Gaussian mixture model fit to the fifth MAD variate after one iteration.

1999 image (reference) using no-change pixels withheld from the normalization procedure are shown in Tables 1 and 2, the former for one iteration of the (i.e., the original) MAD transformation and the latter for iteration to convergence. Results for hypothesis tests for equal means and variances are included in terms of *P*-values. The *P*-values for tests for equal means and variances of the reference and normalized images are in both cases satisfactory, indicating that in this case a single iteration suffices for adequate identification of invariant pixels. (Given the observed value of a test statistic, the *P*-value is the lowest level of significance at which the null hypothesis could have been rejected, i.e., values close to one strongly support the null hypothesis of equal means and variances. Both the paired *t*-tests and the *F*-tests (for the hypotheses of equal means respectively variances after normalization) are so-called two-sided or two-tailed tests. In general we want paired *t*-test values close to zero (the closer the better) and *F*-test values close to one (the closer the better). In both cases we want *P*-values close to one (again, the closer the better). We accept the hypothesis of equal means or equal variances for *P*-values above some pre-set value which is traditionally often set to 5%.)

5.2. Iran

In contrast to the Morocco scene, the ASTER images over Esfahan (see, e.g., Fig. 2) exhibit a very large proportion of change. The statistics for the radiometric normalization of the May 2005 ASTER image (target) to the July 2001 image (reference) using withheld no-change pixels are shown in

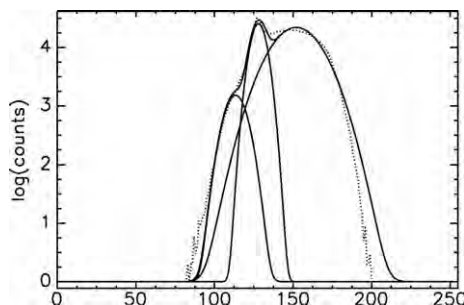


Fig. 9. Gaussian mixture model fit to the fifth MAD variate after iteration to convergence.

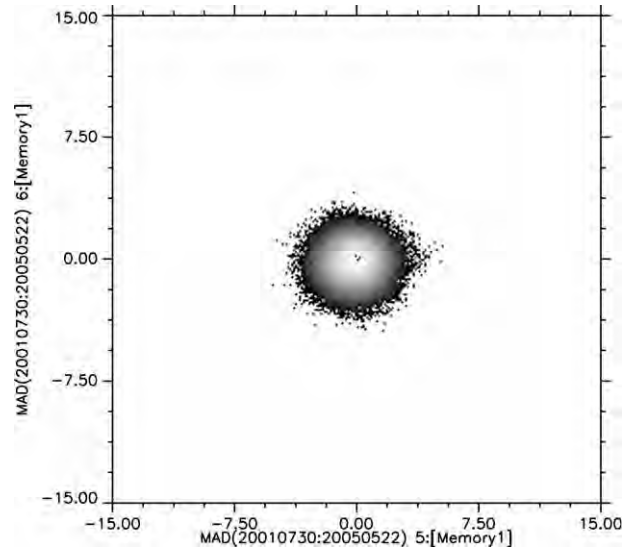


Fig. 10. Scatterplot of MAD variate 6 vs. MAD variate 5 for one iteration.

Tables 3 and 4, the former again for one iteration of the MAD transformation and the latter for iteration to convergence.

Again, the *P*-values for band-wise tests for equal means are acceptable in both cases, however the *P*-values for tests for equal variances of the reference and normalized images are satisfactory only for the IR-MAD result. Figs. 6 and 7 show typical regression lines obtained in the two cases. It is evident that invariant pixels are not properly discriminated after only one application of the MAD transformation, but that the iterative re-weighting scheme succeeds extremely well in isolating them.

This behavior is confirmed in Figs. 8–11. Fig. 8 shows a 3-Gaussian mixture model fit on a logarithmic scale to one of the MAD variates after one iteration, where the central peak corresponds to no-change observations. The Gaussian density

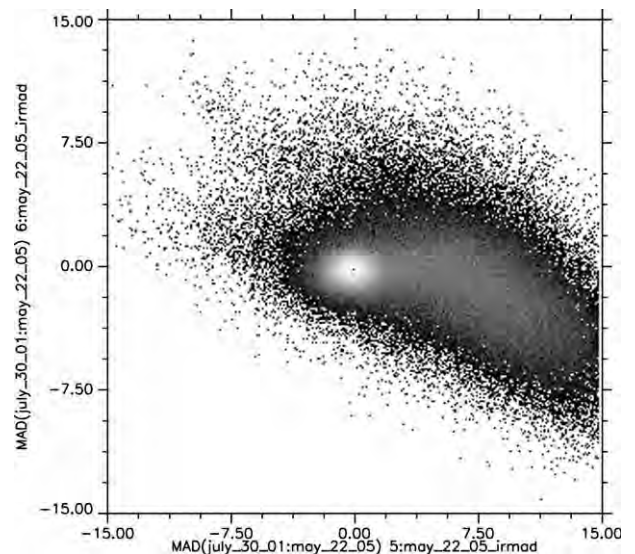


Fig. 11. As Fig. 10 after iteration to convergence.

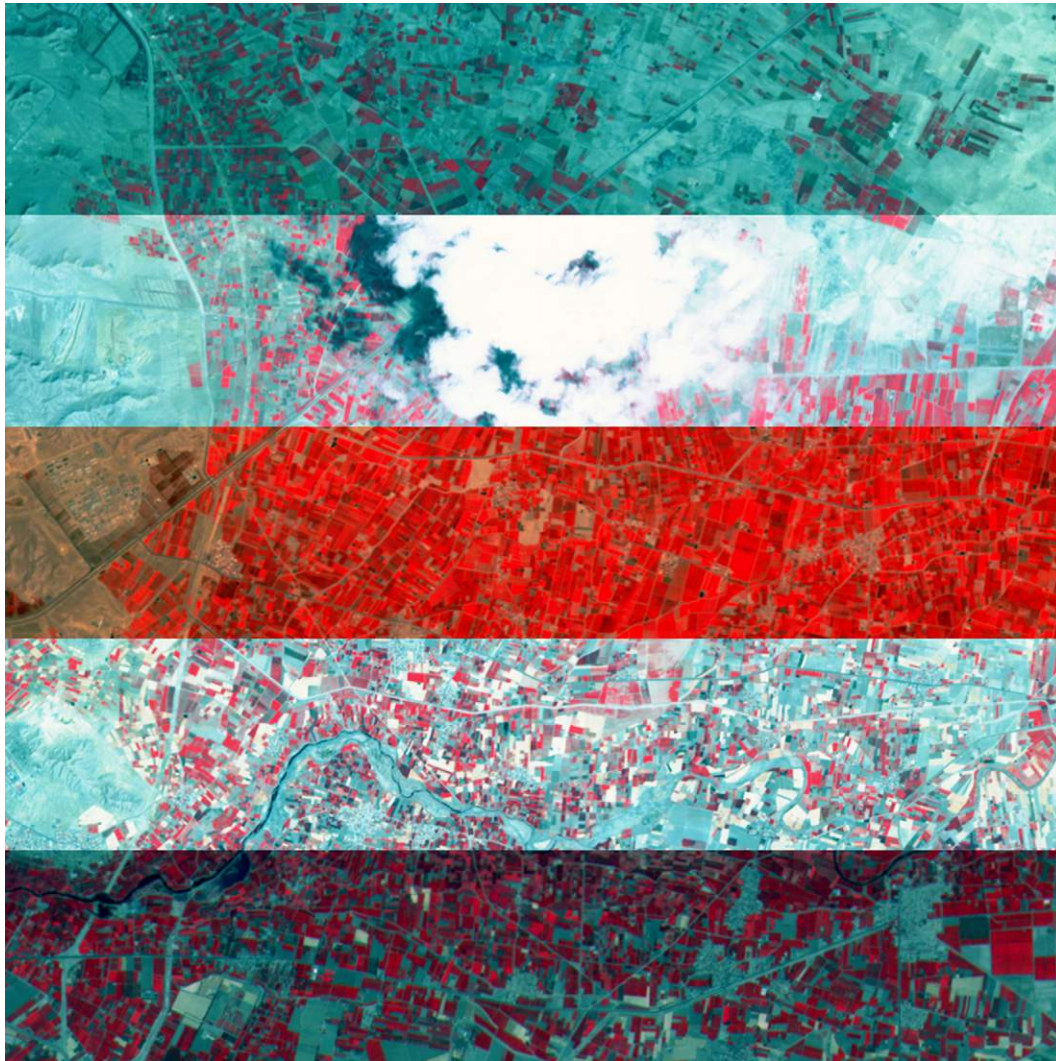


Fig. 12. Mosaic of 5 ASTER images over Iran prior to radiometric normalization, RGB color composite equalization stretch: band 1 (blue), band 2 (green), band 3N (red). The acquisition dates are, from top to bottom: March 11 2002, April 9 2001, May 22 2005, July 30, 2001 and September 11 2005.

parameters were estimated with the expectation maximization algorithm (Redner and Walker, 1984), see Canty and Nielsen (2006) for a detailed discussion. The same fit after iteration to convergence of the MAD transformation is shown in Fig. 9. There, a much better discrimination of the (relatively few) invariant pixels is clearly evident. The scatterplots in Figs. 10 and 11 indicate the same effect in the plane of MAD variates 5 and 6. The no-change observations are the “hot spots” near the origin.

Horizontal mosaics of all 5 ASTER images before and after radiometric normalization to the July 30 2001 image are displayed in Figs. 12 and 13. The selected invariant pixels for each of the other four images, paired with the July 30 2001 scene, were used for the normalization. In Fig. 12 no gain corrections have been applied (this is unnecessary due to the linear invariance of the MAD transformation), partly accounting for the large intensity differences. For example, in the May 22, 2005 image, the 3N band gain was “normal” as opposed to “high” for the reference July 30, 2001 image (see Abrams et al. (1999) for a definition of the

ASTER sensor gain categories). The remaining differences are due to varying atmospheric and solar illumination conditions as well as to true surface reflectance changes.

The success of the normalization is particularly evident on the left hand edge of Fig. 13, where a coherent arid area with no vegetation change can be followed from top to bottom of the mosaic with no discernible transitions, although very significant and real changes occur elsewhere in the scene. Note that the substantial cloud cover in the April 9, 2001 image has no effect on the quality of its normalization as all cloud and cloud shadow pixels are correctly identified as change.

### 5.3. Germany

The results of normalization of the May 25 and August 29, 2001 Jülich LANDSAT ETM+ images to the June 26, 2001 image were comparable to those for the ASTER images: A single application of the MAD transformation did not give acceptable *P*-values, whereas the IR-MAD method did. As an

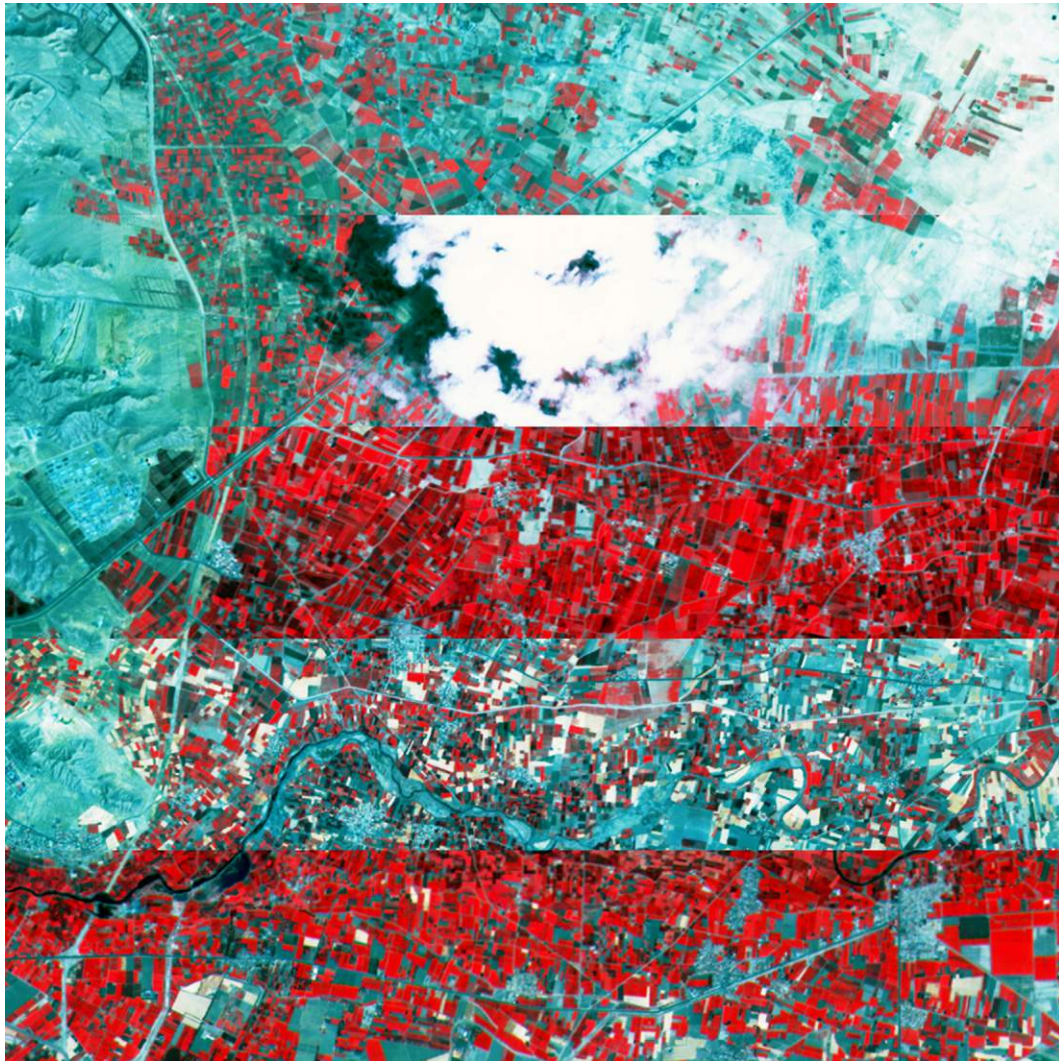


Fig. 13. As Fig. 12 after IR-MAD radiometric normalization.

example, Table 5 gives the comparison statistics for the August 29 (target) to June 26 (reference) normalization with the IR-MAD procedure.

Fig. 14 displays a color composite of three IR-MAD variates for the Jülich scene showing changes between June 26 and August 29,

Table 5  
Comparison of means and variances for 5850 hold-out test pixels, with paired *t*-tests and *t*-tests for equal means and variances of the Jülich images from June 26 and August 29 2001, normalization using the IR-MAD transformation

	Band 1	Band 2	Band 3	Band 4	Band 5	Band 7
Target mean	67.388	51.217	45.538	64.086	60.583	37.262
Ref. mean	76.4787	58.517	49.235	85.411	72.198	41.288
Norm. mean	76.449	58.502	49.228	85.368	72.174	41.281
<i>t</i> -stat.	0.851	0.480	0.181	0.558	0.508	0.230
<i>P</i> -value	0.394	0.631	0.856	0.576	0.611	0.817
Target var.	62.08	79.80	165.64	262.50	206.82	178.03
Ref. var.	82.11	111.78	262.76	531.31	329.05	312.81
Norm. var.	81.92	111.96	263.92	542.57	331.76	316.15
<i>F</i> -stat.	1.002	1.001	1.004	1.021	1.008	1.010
<i>P</i> -value	0.931	0.951	0.866	0.422	0.753	0.6853

2001. The intensities are stretched over  $\pm 8$  standard deviations of the no-change pixels as determined from Gaussian mixture model fits like that shown in Fig. 9. (This very large dynamic range is indicative of the sensitivity of IR-MAD.) Middle gray indicates no change, while colored pixels signify change. Regions of prominent change are cultivated fields (principally due to maturing corn and sugar beet crops as well as grain harvesting), land reclamation in the wake of open pit mining, and mining activities themselves. The change categories are unrelated to one another and the MAD method associates them with different (orthogonal) MAD variates, as indicated by the coloring.

The invariant pixels found by setting a 95% threshold on the no-change probability (see Section 2) are indicated in yellow. They are concentrated in areas of urban settlement, small water bodies (flooded gravel pits) and mixed forest, where there is expected to be little or no change in canopy reflectance between June and August. There is, however, a conspicuous absence of no-change pixels in the “Eifel Forest” area at the bottom of the image, probably due to the more extreme bidirectional reflectance effects of local solar incidence angles associated

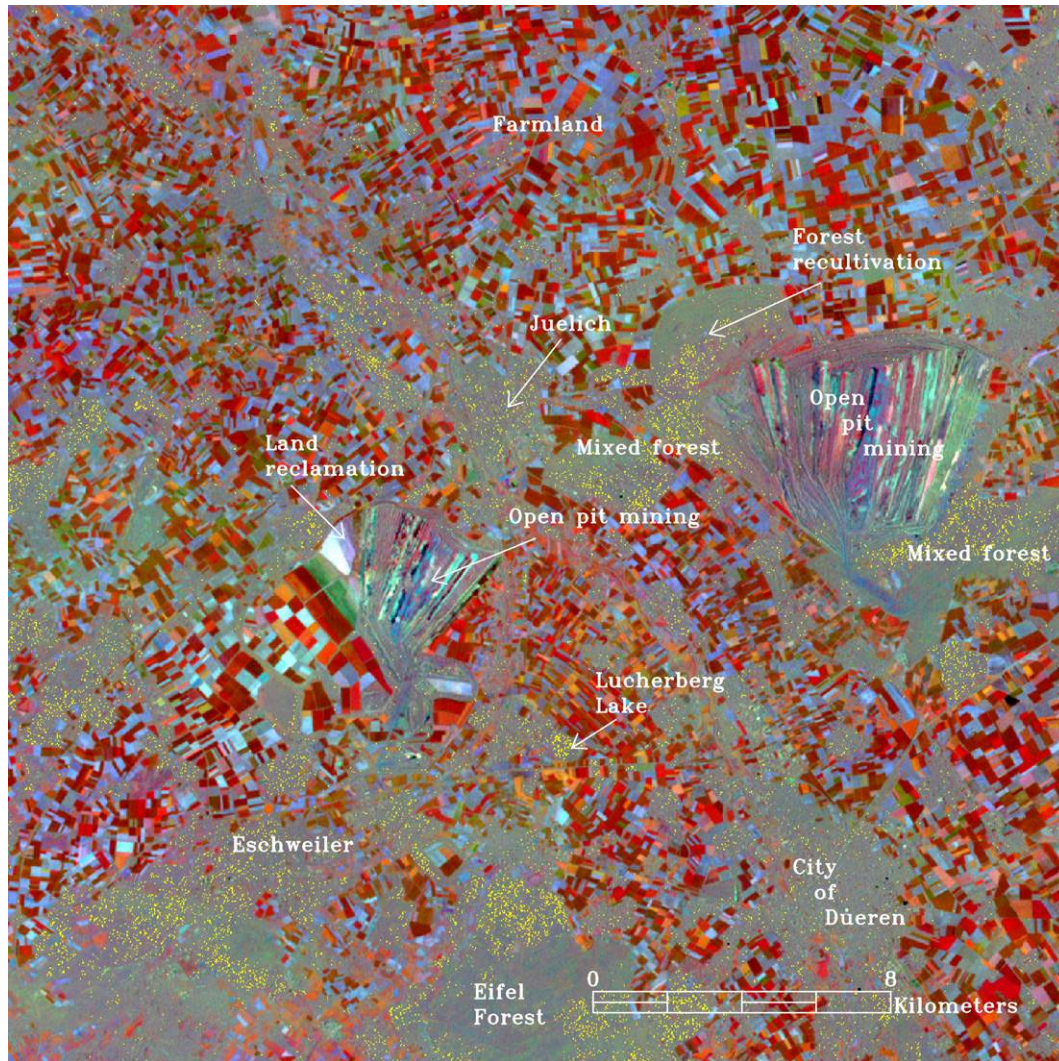


Fig. 14. RGB color composite of MAD variates 4 (blue), 5 (green) and 6 (red) for iteratively re-weighted MAD applied to the June 26 and August 29 Jülich images. The bands are stretched to  $\pm 8$  standard deviations of the no-change pixels. Middle gray signifies no change. The 11,700 invariant pixels determined for radiometric normalization with the IR-MAD procedure are shown in yellow.

with the hilly topography of that region. If we define the no-change fraction to be the relative area under the central Gaussian of a three-Gaussian fit to an iterated MAD variate (see for example Fig. 9), then the no-change fraction for Fig. 14 is roughly 30%. Only a small proportion of these (the yellow pixels) are used for normalization.

## 6. Discussion and conclusions

The ability of the iteratively reweighted MAD transformation to select invariant pixels in the presence of a high background of change has been demonstrated convincingly with partially simulated data. The statistical tests on holdout invariant pixels, as shown in Tables 1–5, also indicate convincingly that, for real bitemporal scenes exhibiting a large amount of change, the MAD method for performing radiometric normalization is significantly improved by using iterative re-weighting.

As opposed to the change pixels, the invariant pixels determined by the MAD or IR-MAD methods appear to be

spatially incoherent, see Fig. 14. This is due to the fact that a high threshold has been used to select them. The no-change observations on the whole will have, as we have indicated, a multivariate normal distribution. The decision threshold  $t$  applied to the no-change probabilities, Eq. (6), will therefore have the effect of truncating most of the distribution and selecting only no-change observations with very small chi-square values. These will appear to be distributed randomly over the areas of no change.

Further insight into the IR-MAD method is provided in Tables 6–8, which illustrate the effect of iteration on the covariance structure of the MAD variates in the case of the Jülich data. Whereas, according to Eq. (3), after a single application of the MAD transformation the variance–covariance matrix of the MAD variates is necessarily diagonal (Table 6), after iteration this is no longer the case (Table 7). However, the variance–covariance matrix of MAD variates based on the weighted (i.e., the no-change) observations is diagonal, as shown in Table 8. The diagonal values in Table 7 are much larger than the weighted

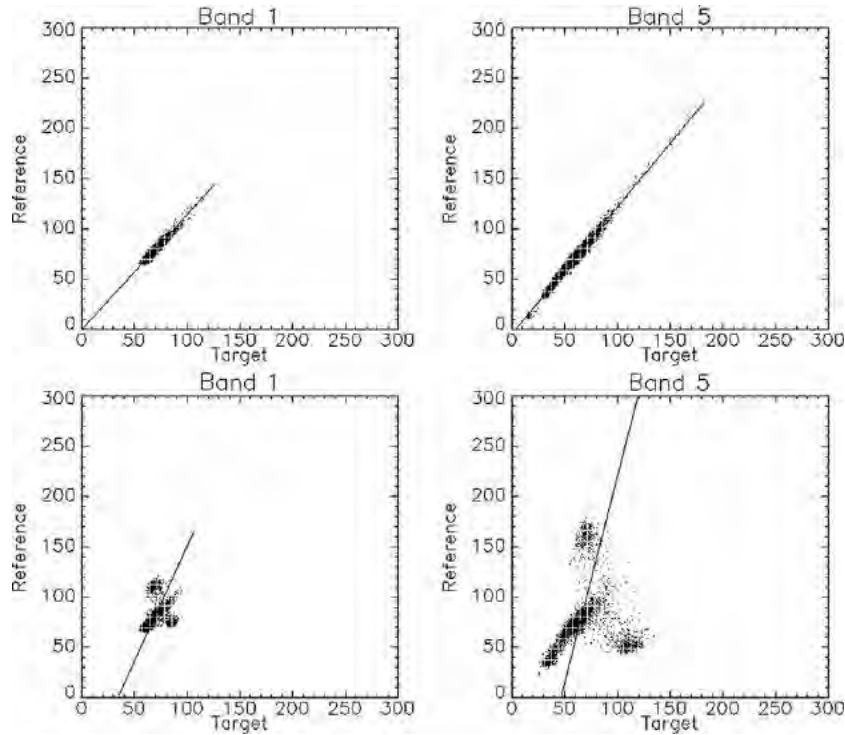


Fig. 15. Regressions on spectral bands 1 and 5 of the Jülich images after iteration of the MAD transformation to convergence. Top row: June 26 (reference) vs. August 29 (target), bottom row: May 25 (reference) vs. August 29 (target). Only pixels identified as unchanged are plotted.

values of Table 8 and the off diagonal values are nonzero. They are in fact determined primarily by the change observations.

The normalization results for the May 2005 and July 2001 ASTER images discussed in Section 5.2 remain essentially unchanged if a 9-band spatial subset consisting of <10% of the area of the first image is copied to the second image, whereas copying a larger subset will cause the IR-MAD algorithm to converge to the artificial no-change pixels, as was discussed in Section 4. The algorithm is evidently “attracted” to those pixels which are most strongly linearly correlated in the two acquisitions and which are in the majority, in the former case to the genuine no-change pixels at the expense of the artificial ones. It may be expected that there will be a lower limit on the number of truly invariant pixels in a bitemporal scene below which the IR-MAD algorithm will fail, in the sense that it will converge to noninvariant observations. This effect, which might be expected to be highly scene-dependent, is illustrated in Fig. 15 with the Jülich data. The regression lines for June 26 vs. August 29 (top row) are satisfactory and correspond to the results of Table 5. Those for May 25 vs. August 29 (bottom row)

are obviously unsuitable for radiometric normalization. The seasonal differences between the two acquisitions (spring to late summer) are evidently too great. Placing a more severe threshold for identifying no-change observations did not remove the two “satellite” clusters in the scatterplots shown. On the other hand, normalization of the May 25 and August 29 images to the June 26 image posed no difficulty.

Quite generally we can conclude that, for multi-temporal images in which change is dominant, the IR-MAD transformation succeeds in identifying suitable invariant pixels for radiometric normalization where the single MAD transformation performs less successfully. For  $1000 \times 1000 \times 6$ -dimensional image arrays (e.g., the Jülich images), a single iteration of the MAD transformation, including determination of the weights for the next iteration, requires about 10 seconds CPU time on a 2.4 GHz Pentium 4 PC running IDL 6.2 under Windows XP. Convergence is usually satisfactory within 20–30 iterations. For very large spatial/spectral subsets (for example LANDSAT full scenes) it is often sufficient to run the transformation on a smaller, representative spatial subset in order to determine a

Table 6  
Variance–covariance matrix of the MAD variates for the Jülich images from June 26 and August 29 2001 after one application of the MAD transformation

1.54	0.00	0.00	0.00	0.00	0.00
0.00	1.25	0.00	0.00	0.00	0.00
0.00	0.00	0.89	0.00	0.00	0.00
0.00	0.00	0.00	0.61	0.00	0.00
0.00	0.00	0.00	0.00	0.52	0.00
0.00	0.00	0.00	0.00	0.00	0.30

Table 7  
As Table 6 after iteration to convergence

5.85	3.87	−1.53	0.62	−2.49	−3.19
3.87	9.17	−3.61	−0.01	−2.70	−3.26
−1.53	−3.61	5.51	−0.22	1.40	1.08
0.62	−0.01	−0.22	1.93	−0.73	−1.00
−2.49	−2.70	1.40	−0.73	2.40	2.53
−3.19	−3.26	1.08	−1.00	2.53	3.50

Table 8  
Weighted variance–covariance matrix of the MAD variates after iteration to convergence

0.664	0.000	0.000	0.000	0.000	0.000
0.000	0.507	0.000	0.000	0.000	0.000
0.000	0.000	0.295	0.000	0.000	0.000
0.000	0.000	0.000	0.122	0.000	0.000
0.000	0.000	0.000	0.000	0.043	0.000
0.000	0.000	0.000	0.000	0.000	0.021

satisfactory regression for radiometric normalization (Schroeder et al., 2006). For mosaicing applications, only overlapping areas can be used, again corresponding to relatively small array dimensions. The iterated MAD scheme thus provides a fast, reliable and robust method for automatic radiometric normalization of multispectral image time series in most situations that one might expect to encounter in practice.

### Acknowledgment

The work presented here was carried out partly within the Global Monitoring for Security and Stability (GMOSS) Network of Excellence of the European Commission.

### References

- Abrams, M., Hook, S., & Ramachandran, B. (1999). ASTER user handbook. *Technical report*. Jet Propulsion Laboratory and California Institute of Technology.
- Aiazzi, B., Alparone, L., Baronti, S., & Garzelli, A. (2002). Context-driven fusion of high spatial and spectral resolution images based on oversampled multiresolution analysis. *IEEE Transactions on Geoscience and Remote Sensing*, 40(10), 2300–2312.
- Canty, M. J. (2007). *Image analysis, classification, and change detection in remote sensing, with algorithms for ENVI/IDL*. Taylor and Francis.
- Canty, M. J., & Nielsen, A. A. (2006). Visualization and unsupervised classification of changes in multispectral satellite imagery. *International Journal of Remote Sensing*, 27(18), 3961–3975. Internet <http://www.imm.dtu.dk/pubdb/p.php?3389>
- Canty, M. J., Nielsen, A. A., & Schmidt, M. (2004). Automatic radiometric normalization of multitemporal satellite imagery. *Remote Sensing of Environment*, 91(3–4), 441–451. Internet <http://www.imm.dtu.dk/pubdb/p.php?2815>
- Coppin, P., Jonckheere, I., Nackaerts, K., & Muys, B. (2004). Digital change detection methods in ecosystem monitoring: A review. *International Journal of Remote Sensing*, 25(9), 1565–1596.
- Du, Y., Teillet, P. M., & Cihlar, J. (2002). Radiometric normalization of multitemporal high resolution images with quality control for land cover change detection. *Remote Sensing of Environment*, 82, 123–134.
- Furby, S. L., & Campbell, N. A. (2001). Calibrating images from different dates to like-valued counts. *Remote Sensing of Environment*, 77, 186–196.
- Hall, F. G., Strebel, D. E., Nickeson, J. E., & Goetz, S. J. (1991). Radiometric rectification: toward a common radiometric response among multitemporal, multisensor images. *Remote Sensing of Environment*, 35, 11–27.
- Hotelling, H. (1936). Relations between two sets of variates. *Biometrika*, 28, 321–377.
- Li, H., Manjunath, B. S., & Mitra, S. K. (1995). A contour-based approach to multisensor image registration. *IEEE Transactions on Image Processing*, 4(3), 320–334.
- Moran, M. S., Jackson, R. D., Slater, P. N., & Teillet, P. M. (1992). Evaluation of simplified procedures for retrieval of land surface reflectance factors from satellite sensor output. *Remote Sensing of Environment*, 41, 160–184.
- Nielsen, A. A. (2007). The regularized iteratively reweighted MAD method for change detection in multi- and hyperspectral data. *IEEE Transactions on Image Processing*, 16(2), 463–478. Internet <http://www.imm.dtu.dk/pubdb/p.php?4695>
- Nielsen, A. A., Conradsen, K., & Simpson, J. J. (1998). Multivariate alteration detection (MAD) and MAF post-processing in multispectral, bitemporal image data: New approaches to change detection studies. *Remote Sensing of Environment*, 64, 1–19. Internet <http://www.imm.dtu.dk/pubdb/p.php?1220>
- Radke, R. J., Andra, S., Al-Kofahi, O., & Roysam, B. (2005). Image change detection algorithms: A systematic survey. *IEEE Transactions on Image Processing*, 14(4), 294–307.
- Redner, R. A., & Walker, H. F. (1984). Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review*, 26(2), 195–239.
- Schmidt, M., King, E., & McVicar, T. R. (2005). Towards an internally consistent calibration for 11 AVHRR instruments in a 20-year time series of satellite data. In W. P. Menzel & T. Iwasaki (Eds.), *Proc. 4th International Asia-Pacific Remote Sensing Symposium, Honolulu* (pp. 5658). Bellingham, Wash.. SPIE.
- Schott, J. R., Salvaggio, C., & Volchok, W. J. (1988). Radiometric scene normalization using pseudo-invariant features. *Remote Sensing of Environment*, 26, 1–16.
- Schroeder, T. A., Cohen, W. B., Song, C., Canty, M. J., & Zhiqiang, Y. (2006). Radiometric calibration of Landsat data for characterization of early successional forest patterns in western Oregon. *Remote Sensing of Environment*, 103(1), 16–26.
- Singh, A. (1989). Digital change detection techniques using remotely-sensed data. *International Journal of Remote Sensing*, 10(6), 989–1002.
- Wiemker, R. (1997). An iterative spectral-spatial Bayesian labeling approach for unsupervised robust change detection on remotely sensed multispectral imagery. *Proceedings of the 7th International Conference on Computer Analysis of Images and Patterns, vol. 1296* (pp. 263–370). LCNS.
- Yang, X., & Lo, C. P. (2000). Relative radiometric normalization performance for change detection from multi-date satellite images. *Photogrammetric Engineering and Remote Sensing*, 66, 967–980.

ANEXO 5.2. RELLENO DE VACÍOS CAUSADOS POR MÁSCARA  
DE NUBES Y SOMBRAS (MARTINUZZI ET AL. 2007)

---



United States  
Department of  
Agriculture

Forest Service

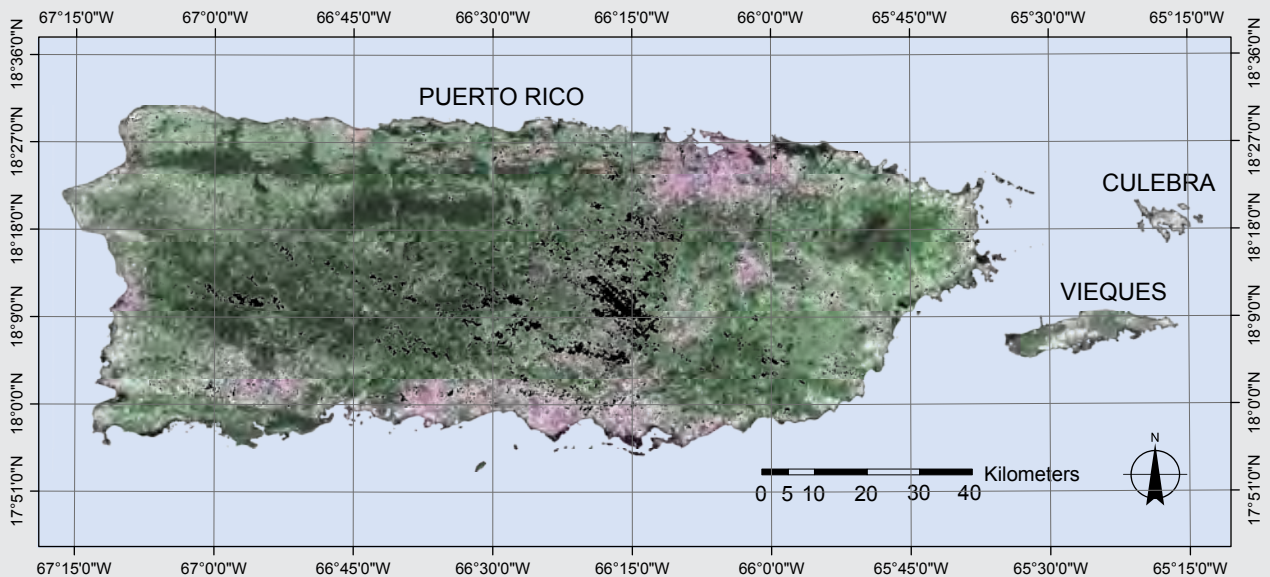
International Institute  
of Tropical Forestry

General Technical  
Report  
IITF-GTR-32  
February 2007



# Creating Cloud-Free Landsat ETM+ Data Sets in Tropical Landscapes: Cloud and Cloud-Shadow Removal

Sebastián Martinuzzi, William A. Gould, and Olga M. Ramos González





The **Forest Service** of the U.S. Department of Agriculture is dedicated to the principle of multiple use management of the Nation's forest resources for sustained yields of wood, water, forage, wildlife, and recreation. Through forestry research, cooperation with the States and private forest owners, and management of the National Forests and National Grasslands, it strives—as directed by Congress—to provide increasingly greater service to a growing Nation.

The U.S. Department of Agriculture (USDA) prohibits discrimination in all its programs and activities on the basis of race, color, national origin, age, disability, and where applicable, sex, marital status, familial status, parental status, religion, sexual orientation, genetic information, political beliefs, reprisal, or because all or part of an individual's income is derived from any public assistance program. (Not all prohibited bases apply to all programs.) Persons with disabilities who require alternative means for communication of program information (Braille, large print, audiotape, etc.) should contact USDA's TARGET Center at (202) 720-2600 (voice and TDD). To file a complaint of discrimination, write USDA, Director, Office of Civil Rights, 1400 Independence Avenue, SW, Washington, DC 20250-9410 or call (800) 795-3272 (voice) or (202) 720-6382 (TDD). USDA is an equal opportunity provider and employer.

## **Authors**

**Sebastián Martinuzzi** is an ecologist, Laboratorio de Investigación de Sistemas Ecológicos y Ambientales, Universidad Nacional de La Plata, Diagonal 113 s/n Escuela Superior de Bosques, CP 1900, Argentina, and visiting scholar at the International Institute of Tropical Forestry. **William A. Gould** is a research ecologist, and **Olga M. Ramos González** is a geographic information system analyst, U.S. Department of Agriculture, Forest Service, International Institute of Tropical Forestry, Jardín Botánico Sur, 1201 Calle Ceiba, Río Piedras, PR 00926-1119.

## **Abstract**

**Martinuzzi, Sebastián; Gould, William A.; Ramos González, Olga M. 2006.**

Creating cloud-free Landsat ETM+ data sets in tropical landscapes: cloud and cloud-shadow removal. Gen. Tech. Rep. IITF-32. Rio Piedras, PR: U.S. Department of Agriculture, Forest Service, International Institute of Tropical Forestry. 12 p.

Clouds and cloud shadows are common features of visible and infrared remotely-sensed images collected from many parts of the world, particularly in humid and tropical regions. We have developed a simple and semiautomated method to mask clouds and shadows in Landsat ETM+ imagery, and have developed a recent cloud-free composite of multitemporal images for Puerto Rico and its adjacent islands that can be used for a variety of landscape analyses. Our assumption is that if clouds and shadows can be identified in a reference image, they can be replaced with data from other dates. We created cloud masks by using Landsat ETM+ band 1 (blue) and thermal band 6.1. Additionally, Landsat ETM+ band 4 (near infrared) and parameters of sun angle, topography, and cloud-shadow projection were used for directing and masking shadows. This methodology was applied to a set of 18 images from 1999 to 2003 to develop an island-wide image that is 96.5 percent cloud free. We considered the seasonality of the imagery when selecting reference images and building the mosaic in order to minimize variation in reflectance related to dry or wet season canopy characteristics. We developed a higher resolution data set by merging the 15-m resolution panchromatic band with the 30-m resolution Landsat ETM+ data. The methodology developed is simple and straightforward to use wherever obtaining cloud-free image data sets is desirable and can be integrated into other efforts that demand an accurate method for the identification of clouds and shadows.

Keywords: Remote sensing, clouds, shadows, mask, Landsat ETM+, Puerto Rico.



## **Introduction**

Clouds are common features of visible and infrared remotely-sensed images collected from many tropical, humid, mountainous, and coastal regions of the world. Puerto Rico is a mountainous tropical island with pressing needs for readily applicable methods to develop Landsat ETM+ cloud-free scenes for analyses of current land cover and land-cover change. Landsat data has been used in Puerto Rico for high temporal resolution analyses of land-cover change related to urban expansion and regeneration of forests over abandoned agricultural land (Grau et al. 2003; Helmer 2004; Helmer and Ruefenacht 2005; Lopez et al. 2001; Thomlinson et al. 1996; Ramos-Gonzalez 2001), estimation of forested area, identification of forests types (Helmer et al. 2002), and current land-cover classification, a key component for mapping habitats and biodiversity as part of the Puerto Rico Gap Analysis Program (Gould 2005). Cloud-free data acquisition for a single date is difficult or impossible for the entire island, particularly for higher elevations where a high percentage of the protected forest resources are located. In the past, the continued presence of clouds in imagery has reduced the amount of island-wide data available, and as a result has restricted analyses to those years for which low cloud cover images exist (1985, 1991, and 1992). Cloud-free image data sets are useful both for the creation of composites and for comparison of temporal variation in spectral signals. In the creation of composite imagery, data can be radiometrically corrected, mosaicked, and analyzed as a single image for classification of general land cover types, or scenes can be classified individually and mosaicked postclassification.

The Puerto Rican landscape represents a challenge for clouds and shadow identification in Landsat ETM+ images owing to the complexity and heterogeneity of land cover (including different forest types and a variety of high-reflectance surfaces such as urban, barren, quarries, rocks, salt mining, sand, pastures and agricultural lands), and topography. In the visible and near-infrared spectral regions of Landsat ETM+ images, clouds are generally characterized by reflectance higher than the underlying (land or ocean) surface. However, under certain surface conditions, such as urban/built-up areas (as for ice and snow in temperate and cold regions), the reflectance difference between cloud and underlying surface can be insufficiently low or even reversed (Li et al. 1999, Schröder et al. 2002). Optical remote sensing images also have the problem of cloud shadows. The spectral irradiance reaching the ground in cloud-shadow areas is both reduced in intensity and changed in its spectral properties (Choi and Milton 1999), similar to that of topographic shadows. The presence of two additional thermal bands in Landsat imagery

(bands 6.1 and 6.2) allows for the analysis of the surface in terms of ground or canopy surface temperatures (Melesse and Jordan 2002). In these bands, clouds, tree foliage, and damp ground appear in dark tones (cooler), whereas grassland and other low-lying vegetation, dry ground, concrete, and asphalt appear in brighter tones (warmer) (Jensen 2000).

The development of cloud-free composites involves both the identification of cloud features as well as distinguishing cloud shadows from topographic shadows. Various techniques have been developed to minimize the effect of clouds and cloud-shadows (Ackerman et al. 1998, Choi and Milton 1999, Helmer and Ruefenacht 2005, Li et al. 1999, Liew et al. 1998, Milton and Choi 1999, Schröder et al. 2002, Song and Civco 2002). We used optical and thermal spectral characteristics as well as geographic position to identify cloud and cloud shadows. Analysis of spectral reflectance in different bands is essential to identifying features in the images. Song and Civco (2002) used the brightness value in Landsat TM bands 1 (blue) and 4 (near infrared) to detect clouds and shadows in Madagascar. We added thermal information from Landsat ETM band 6.1 to spectrally separate clouds from similarly responding features, such as urban/built-up lands. Although cloud shadows and topographic shadows are not spectrally distinguishable in optical images, they can be identified by geographic position. The direction of the shadow occurrence relative to the cloud is the same for the entire scene as it depends on the sun illumination angle, a constant for any given scene. The location, form, and dimensions of any cloud shadow depend on the cloud shape and elevation, topography, and the sun illumination angle.

The development of masks for identifying clouds and cloud shadows is our solution to develop a final cloud- and shadow-free Landsat ETM+ composite image. Our principle is to identify clouds and shadows in a reference image and replace them with cloud- and shadow-free parts from other dates. Masks facilitate operations such as subtraction and replacement of specific areas, as well as buffering, intersection, and union.

We have developed simple semiautomated procedures to generate cloud and shadow masks in tropical regions for use in the creation of cloud-free composite Landsat ETM+ imagery, by using multiple recent images acquired at different times over the same region. Our goals in developing these methods were to make use of spectral and geographic information to facilitate the identification of cloud cover, separate cloud and topographic shadows, and minimize the needed masked areas.

In addition to the masking method, we have incorporated additional processes for the creation of the final imagery data set. We sought to minimize differences in reflectance owing to atmospheric conditions by correcting for Rayleigh scatter and differences in seasonality of vegetation by using, whenever possible, imagery acquired during the same season. Aside from Rayleigh scatter correction, we have not incorporated any methods to normalize systematic scene-to-scene radiometric variation such as dark object subtraction (Chavez 1996). This would be advisable when classifying a composite image but unnecessary when classifying individual scenes before mosaicking. We also improved our ability to separate subpixel features (for urban or mangrove areas) by enhancing the spatial resolution incorporating the higher (15-m) resolution of the panchromatic band into the multispectral images. We have developed a useful image data set for the islands of Puerto Rico, Vieques, Culebra, and additional small islands within the Puerto Rican Archipelago. The methods to develop this data set will work equally well throughout the Caribbean basin and similar regions, and with Landsat ETM+ imagery acquired over a range of dates.

## **Data and Methods**

Four Landsat ETM+ path-rows (4-47, 4-48, 5-47, and 5-48) are necessary to cover Puerto Rico and its adjacent islands of Vieques and Culebra. Our data set is composed of 18 images acquired between September 1999 and February 2003. During 2002, no images were incorporated owing to persistent dense cloud cover. The complete process leading to the resulting cloud- and cloud-shadow-free composite image can be accomplished in four major steps outlined in our methods, and subsequently applied to our 18 images. These include atmospheric correction, cloud and cloud-shadow masking, resampling and addition of the panchromatic band, and image selection and composition.

### **Atmospheric Correction**

Atmospheric correction is necessary to put multitemporal data on the same radiometric scale, in order to create an image data set with compatible scenes that is useful for classification and monitoring terrestrial surfaces over time (Song et al. 2001). We used a model developed by the Institute for Astronomy and Space Physics IAFE-CONITET to convert radiance to reflectance corrected by Rayleigh scatter (Paolini et al. 2004). The input parameters required for running the model are included in the image header and include gain, bias, sun elevation angle, solar azimuth angle, and date. The model is described by the following formula:

$$\text{Surface reflectance} = \delta \times (L_{\bar{e}} - L_{\text{Rayleigh}}) \times r^2/E_o (\cos/\delta) ,$$

Where

$L_{\bar{e}}$  = at satellite radiance =  $\text{gain}_{\bar{e}} \times \text{DN} \times \text{bias}_{\bar{e}}$ ,

$L_{\text{Rayleigh}}$  = radiance owing to Rayleigh scatter,

$E_o$  = extraterrestrial irradiance,

$r$  = Earth-Sun distance, and

$\delta$  = Sun elevation angle.

Each scene was processed using this formula before proceeding with the cloud and shadow masking.

### Cloud and Cloud-Shadow Masking

This involves two main procedures: (1) identifying clouds and developing a mask by using Landsat ETM+ brightness values (DN) in optical band 1 (blue) and in thermal band 6.1 and (2) developing a final cloud and cloud-shadow mask by using the cloud mask, the brightness values of shadowed areas in band 4 (near infrared), and the geometric relationship between shadow position and the clouds (i.e., parameters of cloud projection) (fig. 1).

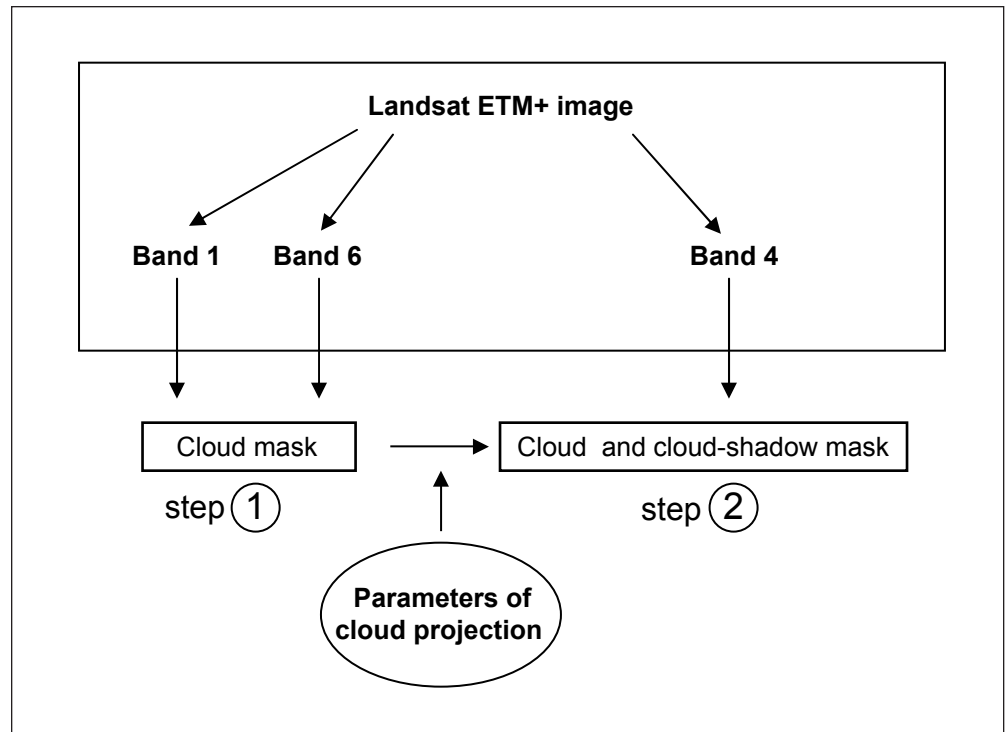


Figure 1—Steps for masking clouds and cloud shadows. In step 1, the range of reflectance values for clouds are determined by visually inspecting the range of cloud types and digital numbers (DNs) in band 1 (blue) and in thermal band 6.1, and then selecting DN's in this range. In step 2, band 4 (near infrared), the geometric relationships of shadows to clouds, and the clouds mask, were used in the identification of cloud shadows.

## Cloud Masking

The spatial resolution of the thermal band 6.1 was resampled from 60- to 30-m pixels, and subsequently bands 1 and 6 were analyzed by using the following procedures:

- Brightness values for clouds were identified by visual analysis. For band 1, DN values between 120 and 255 include clouds as well as urban, barren, quarries, rocks, and sand. For band 6, DN values of 102 to 128 include both clouds and densely forested areas.
- A mask was created for band 1 and for band 6 by using the previous values. A three-pixel buffer was added to incorporate mixed pixels from cloud borders that could not be incorporated by brightness value analysis.
- The intersection of both masks results in a final clouds mask, which was also buffered by three pixels (fig. 2).

## Cloud-Shadow Masking and Obtaining the Final Cloud and Cloud-Shadow Mask

The first step in the creation of a cloud-shadow mask was to estimate the potential shadowed areas in the reference image by projecting the cloud mask on the land, and considering the effects of Sun illumination angle, topography, and cloud elevation (fig. 3). For this, an identifiable feature or point observable in both a cloud and its shadow was located in the image, ideally over low-elevation landscapes where clouds are at the greatest height above land (assuming a relatively constant elevation of clouds affecting the image). The difference in X and Y coordinates for these points is an estimation of the horizontal “shift” between cloud and shadow positions. A mask of potential cloud-shadow areas was created by displacement of the cloud mask based on the horizontal shift. To account for possible variations in the shadowed areas related to variation in cloud elevation and topography, a 10-pixel buffer was added. The resulting mask includes and overestimates all the areas with potential cloud shadows. Within these areas, the brightness values in band 4 were used to differentiate cloud-shadow areas from nonshadowed areas (fig. 2). In our imagery, the DN range 17 to 66 includes cloud shadows, but also some topographic shadows, certain urban areas, and land-water transition zones that are within potential cloud-shadow areas. Finally, cloud-shadow areas were added to the cloud mask to obtain a final cloud and cloud-shadow mask.



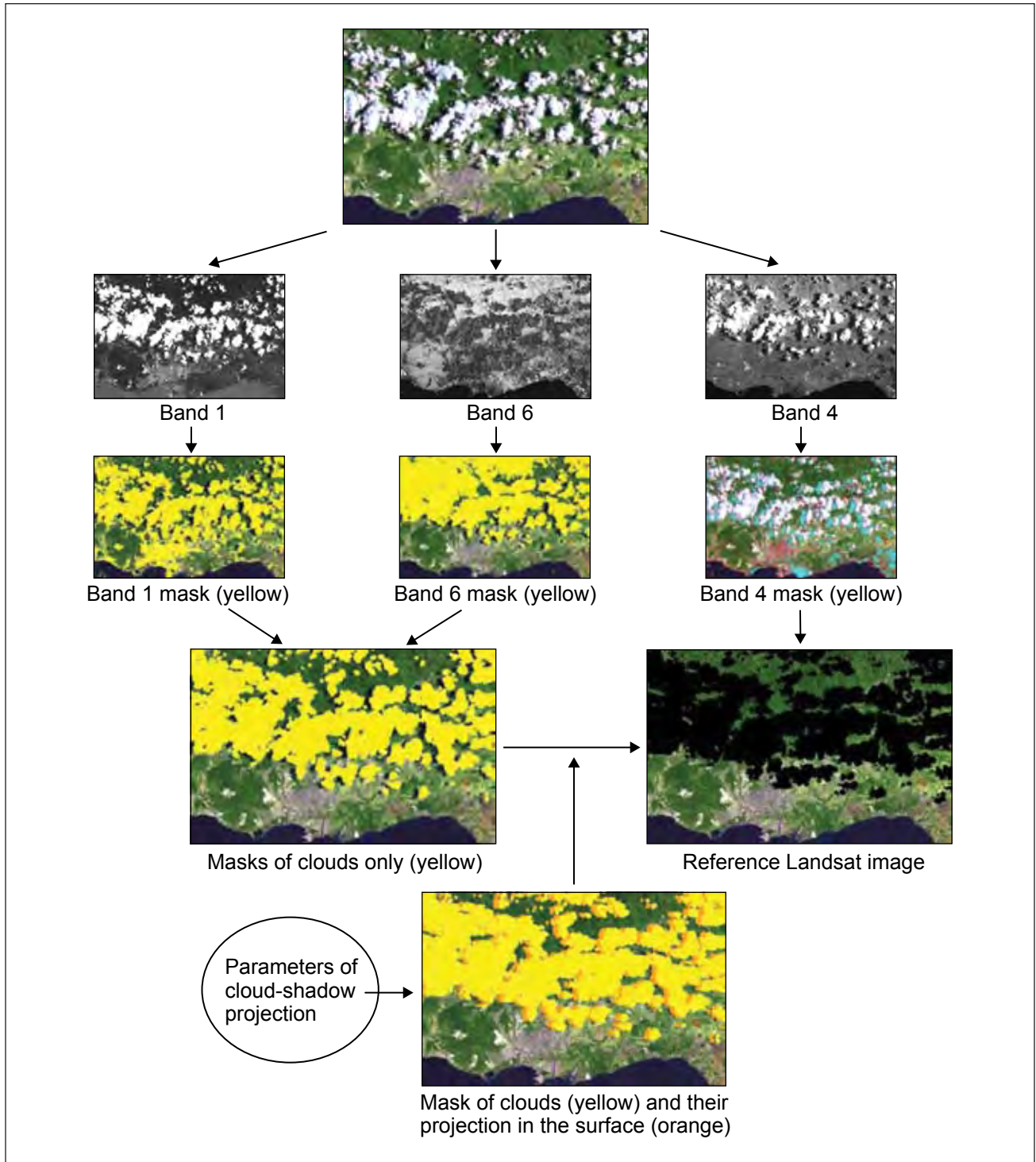


Figure 2—Flow chart illustrating cloud and cloud-shadow masking steps in an area of heterogeneous land cover and high cloud cover. A Landsat ETM+ subset image of the town of Ponce in southern Puerto Rico includes urban areas, cultivated land, pastures, quarries, barren areas, beaches, rivers, clouds, and shadows. The image is shown in band combination R (7) G (5) B (3). Three masks were developed from bands 1, 6, and 4, respectively. Band 1 mask (in yellow) includes clouds and some urban, barren, quarries, rocks, and sand. Band 6 mask (in yellow) includes clouds and some densely forested areas. Band 4 mask (in blue) includes cloud shadows, topographic shadows, certain urban areas, and land-water transition zones. Band 1 and band 6 masks are intersected for a mask of clouds only (yellow). This is joined with a cloud-shadow mask to develop the cloud-free reference image.

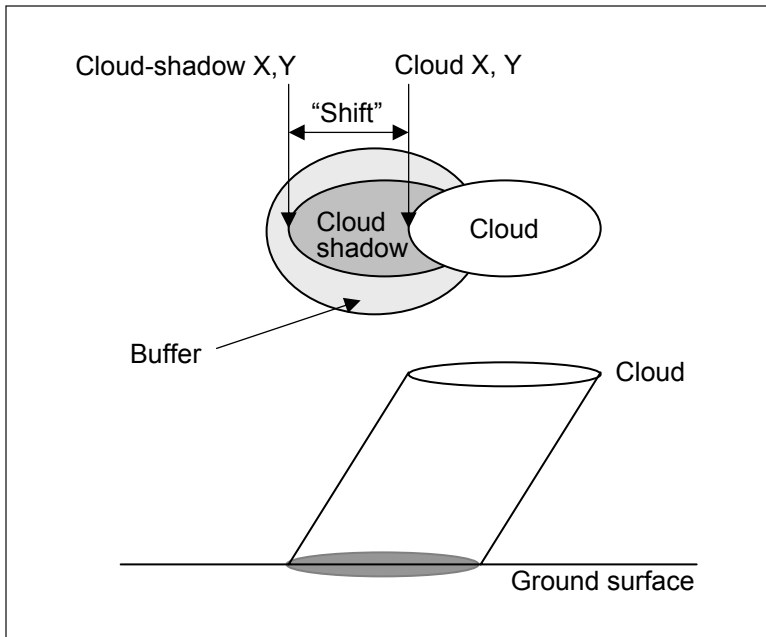


Figure 3—Representation of the geometric shift of cloud shadow plus buffer mask along X and Y axes that is determined by cloud elevation and Sun angle.

## Resolution Enhancement and Image Composition

One of our immediate uses of this composite image data set is to study land use and urban developments. Because much of the spatial variability in developed and undeveloped surfaces is related to roads and buildings that are subpixel features at 30-m resolution, we enhanced the spatial resolution of the images by merging the visible and IR bands (30- by 30-m pixel) with the panchromatic band (15- by 15-m pixel) in all of our scenes by using Principal Component Analysis and Cubic Convolution Resampling in ERDAS Imagine 8.7. For the image composition the most cloud-free scene from each path/row was selected as the reference image, i.e., the top scene for each path/row in the mosaic. Cloud and cloud-shadow areas in reference images were replaced with data from other dates.

## Results

Our set of images ranged from 16.35 to 74.64 percent cloud and cloud-shadow coverage and contained acquisition dates in both the wet and dry seasons in Puerto Rico (table 1). Over half of the scenes had greater than 50 percent cloud cover. Two scenes were from 1999, eight from 2000, seven from 2001, none from 2002, and one from 2003. We qualitatively ranked the images into dry or wet season assisted by monthly rainfall data from different locations in the island, and visual analysis of the images. Seven scenes were from the dry season, nine from the wet season,

**Table 1—Landsat ETM+ imagery data set description, including reference image number, path, row, region covered, acquisition date, season, and percentage of cloud cover**

Reference image	Path	Row	Region	Acquisition date	Season (wet or dry)	Cloud cover
						<i>Percent</i>
1	5	47	NW and Center PR	January 22, 2003	d	27.66
2	5	47	NW and Center PR	September 10, 2000	w	55.80
3	5	47	NW and Center PR	July 27, 2001	w	47.01
4	5	47	NW and Center PR	March 5, 2001	d	53.62
5	5	48	Southwest PR	September 10, 2000	w	52.86
6	5	48	Southwest PR	July 11, 2001	d	39.84
7	5	48	Southwest PR	November 13, 2000	w	16.35
8	4	47	Northeast PR	May 14, 2000	d-w	40.81
9	4	47	Northeast PR	September 17, 1999	w	62.74
10	4	47	Northeast PR	January 25, 2001	w	62.08
11	4	47	Northeast PR	February 26, 2001	d-w	58.43
12	4	47	Northeast PR	March 27, 2000	d	34.65
13	4	47	Northeast PR	August 2, 2000	d	69.00
14	4	47	Northeast PR	January 9, 2001	w	52.24
15	4	48	Southeast PR	September 17, 1999	w	64.40
16	4	48	Southeast PR	August 18, 2000	d	46.76
17	4	48	Southeast PR	January 25, 2001	w	74.64
18	4	48	Southeast PR	March 27, 2000	d	27.39
Final composite						3.50

and two were transitional. The significance of identifying seasonality in the scenes is to maintain, when possible, phenological homogeneity between images composing the mosaic.

For the mosaic, one image from each path/row was selected as the reference image, i.e., the top scene for each path row in the mosaic. To determine which scene would be our reference, we chose scenes low in cloud cover and from the dry season. The primary reference image was path/row 5-47 from January 2003 that covered the north, east, and center of the island and had low cloud cover (27.66 percent). In addition, we used a Landsat 5 image from 1985, the only completely cloud-free image for eastern Puerto Rico at the time, to replace 5600 ha (0.65 percent of the island) of medium- and high-elevation forested areas located within the Caribbean National Forest that had persistent cloud cover in all the recent images. We produced a 15-m Landsat ETM+ composite image data set that is 96.5 percent cloud and cloud-shadow free. The final mosaic (fig. 4) is displayed by using the band combination R (7), G (5), B (3). Forested vegetation shows up as dark green, grassland and agricultural areas as pale green, urban and, dry agricultural areas as violet. Remaining cloud-covered areas (3.5 percent) are in black and will be replaced as cloud-free imagery becomes available.

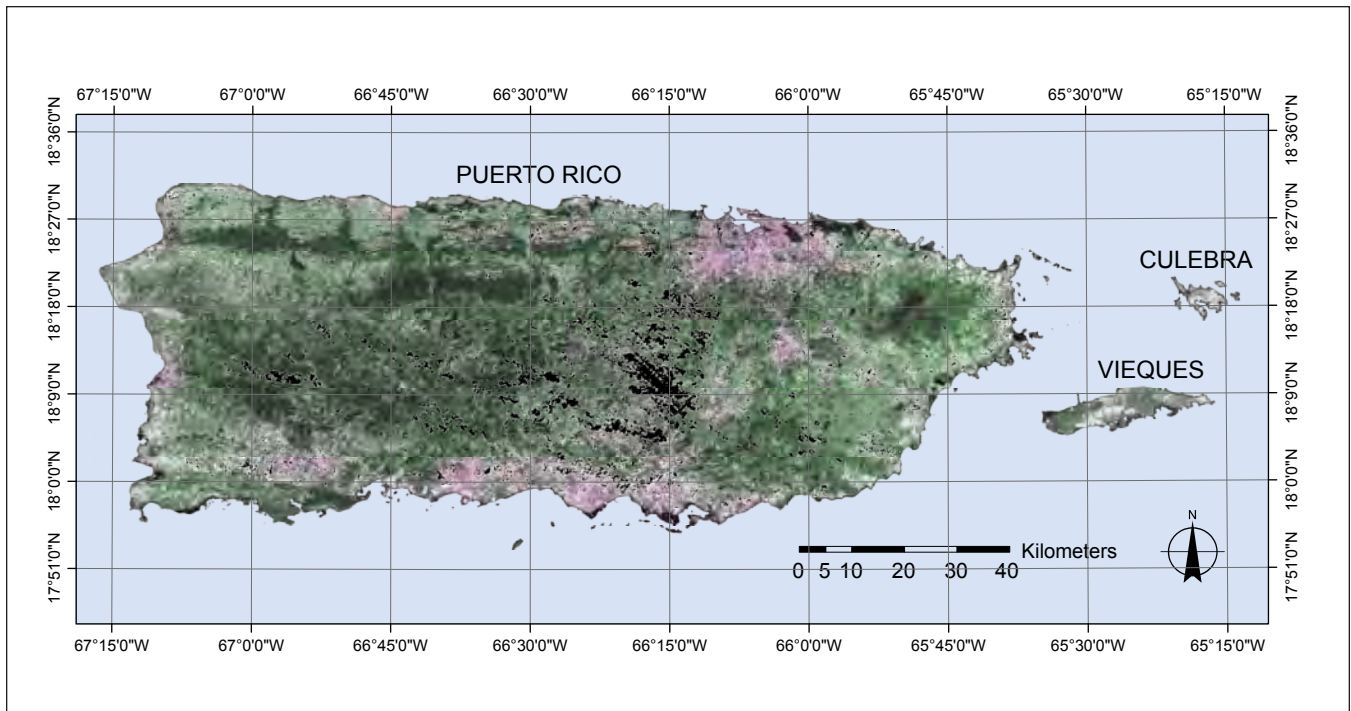


Figure 4—Composite Landsat ETM+ image for Puerto Rico, Vieques, and Culebra, developed from a set of 18 images acquired between 1999 and 2003.

## Discussion

Remote-sensing tools and applications have become deeply integrated in a wide range of theoretical and applied ecological modeling. New problems and questions, new remote-sensing technologies, better computing facilities, and more sophisticated analyses are continually being developed. Advancements in optical remote sensing are particularly useful for characterizing landscapes in fast-changing environments where scientific expertise must keep pace with development and the needs of society. The ability to contribute to these advancements and use them for research and management issues depends on the ability to acquire and build data sets with complete historical and current coverage. It will always be difficult to obtain cloud-free data in cloud-prone environments, but applying simple methods of cloud and cloud-shadow detection, atmospheric correction, and image mosaicking can be beneficial in developing useful remote-sensing data sets.

The major benefit of this methodology is its simplicity and straightforward approach, allowing cloud-free-image analyses by using the most comprehensive image data set for the study area and objectives. The distinctive point of our study is that we semiautomatically derive cloud and cloud-shadow masks from simple parameters. There are a number of instances in which this methodology can prove useful:

- Our method for developing a cloud and cloud-shadow mask is useful for preparing the imagery for pixel-replacement procedures. Innovative strategies for the development of cloud-free imagery use regression trees to predict values underneath clouds and cloud shadows in reference scenes based scenes from other dates (Helmer and Ruefenacht 2005). The input images need to have clouds and cloud shadows already masked, which is originally solved by a combination of unsupervised classification and manual editing.
- Our method increases the functionality of Landsat TM and ETM+ data that have been acquired since 1982, and had been underutilized because of the presence of clouds, and adds to the set of images suitable for conducting modeling and land-cover change studies.
- The method highlights the importance of the thermal band in the Landsat TM and ETM+ sensors for the creation of cloud- and shadow-free images in tropical and other landscapes.
- The method can be incorporated into the research effort of the U.S. Geological Survey (USGS) Landsat 7 Project at the USGS EROS Data Center to increase the utility of the Landsat 7 ETM+ data for images acquired in tropical regions.

The limitations of our methodology include, for the creation of the masks, the inability to discriminate between cloud shadow and topographic shadow within areas of potential shadow occurrence, and in the mosaicking process, issues common to any multiscene or multirate composite, i.e., atmospheric differences and temporal changes related to seasonality of vegetation or land use. We found that the atmospheric correction for Rayleigh scatter eliminated some of the variability between scenes related to atmospheric conditions. The best-case scenario is to identify and minimize likely temporal changes, or develop models to estimate expected spectral responses related to seasonality. In a region like Puerto Rico, with areas of the island having strong vegetation responses to dry and wet seasons, the best-case scenario would be to develop image composites for both the dry and wet seasons.

## Acknowledgments

This research was supported by the International Institute of Tropical Forestry (IITF) and the U.S. Geological Survey, Biological Resources Division, Gap Analysis Program (01-HQPG-003/01-IA-11120101-002). Thanks to Haydee Karszenbaum at the Institute for Astronomy and Space Physics IAFE-CONICET;

Bonnie Ruefenacht, USDA Forest Service, Remote Sensing Applications Center; and Maya Quiñones and Eileen Helmer, USDA Forest Service, IITF. This research was supported by the IITF GIS and Remote Sensing Laboratory. All research at the IITF is done in collaboration with the University of Puerto Rico.

## English Equivalents

<b>When you know:</b>	<b>Multiply by:</b>	<b>To get:</b>
Meters (m)	3.28	Feet
Hectares (ha)	2.47	Acres

## References

- Ackerman, S.A.; Strabala, K.I.; Menzel, W.P.; Frey, R.A.; Moeller, C.C.; Gumley, L.E. 1998.** Discriminating clear sky from clouds with MODIS. *Journal of Geophysical Research*. 103(D24): 32141–32157.
- Chavez, P.S., Jr. 1996.** Image-based atmospheric corrections—revisited and revised. *Photogrammetric Engineering and Remote Sensing*. 62(9): 1025–1036.
- Choi, K-Y.; Milton, E.J. 1999.** A multispectral transform for the suppression of cloud shadows. In: *Proceedings: Fourth international airborne remote sensing conference and exhibition/21<sup>st</sup> Canadian symposium on remote sensing*. Ottawa, Canada: ERIM International Inc. 2: 762–769.
- Grau H.R.; Aide, T.M.; Zimmerman, J.K.; Thomlinson, J.R.; Helmer, E.; Zou, X. 2003.** The ecological consequences of socioeconomic and land-use changes in postagriculture Puerto Rico. *BioScience*. 53(12): 1159–1168.
- Gould, W. 2005.** Puerto Rico Gap Analysis Project Status Report. In: Maxwell, J.M.; Gergely, K.; Aycrigg, J; Beard, D.; Coffey, N.; Sajwaj, T., eds. *Gap Analysis Bulletin No. 13*. Moscow, ID: U.S. Geological Survey, Biological Resources Division, Gap Analysis Program: 68–71.
- Helmer, E.H. 2004.** Forest conservation and land development in Puerto Rico. *Landscape Ecology*. 19: 29–40.
- Helmer, E.H.; Ramos, O.; López, T. del M.; Quiñones, M.; Díaz, W. 2002.** Mapping the forest type and land cover of Puerto Rico, a component of the Caribbean biodiversity hotspot. *Caribbean Journal of Science*. 38(3–4): 165–183.
- Helmer, E.H.; Ruefenacht, B. 2005.** Cloud-free satellite image mosaics with regression trees and histogram matching. *Photogrammetric Engineering and Remote Sensing*. 71: 1079–1089.

- Jensen, J.R. 2000.** Remote sensing of the environment: an Earth resource. Upper Saddle River, NJ: Prentice Hall. 544 p.
- Li, M.; Liew, S.C.; Kwoh, L.K.; Lim, H. 1999.** Improved cloud-free multi-scene mosaics of SPOT images. Proceedings of the 20<sup>th</sup> Asian conference for remote sensing. Hong Kong, China.
- Liew, S.C.; Li, M.; Kwoh, L.K.; Chen, P.; Lim, H. 1998.** Cloud-free multiscene mosaics of SPOT images. Proceedings of the 1998 International Geosciences and Remote Sensing Symposium. 2: 1083–1085.
- López, T. del M.; Aide, T.M.; Thomlinson, J.R. 2001.** Urban expansion and the loss of prime agricultural lands in Puerto Rico. *Ambio*. 30: 49–54.
- Melesse, A.M.; Jordan, J.D. 2002.** A comparison of fuzzy vs. augmented ISODATTA classification algorithms for cloud-shadow discrimination from Landsat images. *Photogrammetric Engineering and Remote Sensing*. 68(9): 905–911.
- Milton, E.J.; Choi, K. 1999.** Cloud shadow suppression using a feature space approach to the identification of virtual endmembers. *RSS '99: from data to information*. Nottingham, United Kingdom: Remote Sensing Society: 751–758.
- Ramos-González, O.M. 2001.** Assessing vegetation and land cover changes in northeastern Puerto Rico: 1978-1995. *Caribbean Journal of Science*. 37: 95–106.
- Song, C.; Woodcock, C.E.; Seto, K.C.; Pax-Lenney, M.; 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.
- Song, M.; Civco, D.L. 2002.** A knowledge-based approach for reducing cloud and shadow. Proceedings of the American Society of Photogrammetry and Remote Sensing annual convention. Washington, DC: American Society of Photogrammetry and Remote Sensing. 7 p.
- Schröder, M.; Bennartz, R.; Schüller, L.; Preusker, R.; Albert, P.; Fischer, J. 2002.** Generating cloud-masks from spatial high resolution observations of clouds using texture and radiance information. *International Journal of Remote Sensing*. 23: 4247–4261.
- Thomlinson, J.R.; Serrano, M.I.; López, T. del M.; Aide, T.M.; Zimmerman, J.K. 1996.** Land-use dynamics in a post agricultural Puerto Rican Landscape (1936–1988). *Biotropica*. 28: 525–536.

**Pacific Northwest Research Station**

<b>Web site</b>	<a href="http://www.fs.fed.us/pnw/">http://www.fs.fed.us/pnw/</a>
<b>Telephone</b>	(503) 808-2592
<b>Publication requests</b>	(503) 808-2138
<b>FAX</b>	(503) 808-2130
<b>E-mail</b>	<a href="mailto:pnw_pnwpubs@fs.fed.us">pnw_pnwpubs@fs.fed.us</a>
<b>Mailing address</b>	Publications Distribution Pacific Northwest Research Station P.O. Box 3890 Portland, OR 97208-3890



---

U.S. Department of Agriculture  
Pacific Northwest Research Station  
333 SW First Avenue  
P.O. Box 3890  
Portland, OR 97208-3890

---

Official Business  
Penalty for Private Use, \$300

ANEXO 5.3. GOOD PRACTICES FOR ESTIMATING AREA AND  
ASSESSING ACCURACY OF LAND CHANGE (OLOFSSON ET AL.  
2014)

---



## Review

## Good practices for estimating area and assessing accuracy of land change



Pontus Olofsson <sup>a,\*</sup>, Giles M. Foody <sup>b</sup>, Martin Herold <sup>c</sup>, Stephen V. Stehman <sup>d</sup>,  
Curtis E. Woodcock <sup>a</sup>, Michael A. Wulder <sup>e</sup>

<sup>a</sup> Department of Earth and Environment, Boston University, 685 Commonwealth Avenue, Boston, MA 02215, USA

<sup>b</sup> School of Geography, University of Nottingham, University Park, Nottingham NG7 2RD, UK

<sup>c</sup> Laboratory of Geo-Information Science and Remote Sensing, Wageningen University, Droevendaalsesteeg 3, 6708 Wageningen, The Netherlands

<sup>d</sup> Department of Forest and Natural Resources Management, State University of New York, 1 Forestry Drive, Syracuse, NY 13210, USA

<sup>e</sup> Canadian Forest Service (Pacific Forestry Centre), Natural Resources Canada, Victoria, BC 12 V8Z 1M5, Canada

## ARTICLE INFO

## Article history:

Received 30 May 2013

Received in revised form 15 January 2014

Accepted 22 February 2014

Available online xxxx

## Keywords:

Accuracy assessment

Sampling design

Response design

Area estimation

Land change

Remote sensing

## ABSTRACT

The remote sensing science and application communities have developed increasingly reliable, consistent, and robust approaches for capturing land dynamics to meet a range of information needs. Statistically robust and transparent approaches for assessing accuracy and estimating area of change are critical to ensure the integrity of land change information. We provide practitioners with a set of “good practice” recommendations for designing and implementing an accuracy assessment of a change map and estimating area based on the reference sample data. The good practice recommendations address the three major components: sampling design, response design and analysis. The primary good practice recommendations for assessing accuracy and estimating area are: (i) implement a probability sampling design that is chosen to achieve the priority objectives of accuracy and area estimation while also satisfying practical constraints such as cost and available sources of reference data; (ii) implement a response design protocol that is based on reference data sources that provide sufficient spatial and temporal representation to accurately label each unit in the sample (i.e., the “reference classification” will be considerably more accurate than the map classification being evaluated); (iii) implement an analysis that is consistent with the sampling design and response design protocols; (iv) summarize the accuracy assessment by reporting the estimated error matrix in terms of proportion of area and estimates of overall accuracy, user's accuracy (or commission error), and producer's accuracy (or omission error); (v) estimate area of classes (e.g., types of change such as wetland loss or types of persistence such as stable forest) based on the reference classification of the sample units; (vi) quantify uncertainty by reporting confidence intervals for accuracy and area parameters; (vii) evaluate variability and potential error in the reference classification; and (viii) document deviations from good practice that may substantially affect the results. An example application is provided to illustrate the recommended process.

© 2014 Elsevier Inc. All rights reserved.

## Contents

1.	Introduction . . . . .	43
1.1.	Good practice recommendations . . . . .	43
1.2.	Context of good practice recommendations . . . . .	44
2.	Sampling design . . . . .	44
2.1.	Choosing the sampling design . . . . .	45
2.1.1.	Strata . . . . .	45
2.1.2.	Cluster sampling . . . . .	46
2.1.3.	Systematic vs. random selection . . . . .	46
2.2.	A recommended good practice sampling design . . . . .	47
3.	Response design . . . . .	47
3.1.	Spatial assessment unit . . . . .	47
3.2.	Sources of reference data . . . . .	47

\* Corresponding author. Tel.: +1 617 353 9734; fax: +1 617 353 8399.

E-mail address: [olofsson@bu.edu](mailto:olofsson@bu.edu) (P. Olofsson).

URL: <http://people.bu.edu/olofsson> (P. Olofsson).

3.3.	Reference labeling protocol . . . . .	49
3.4.	Defining agreement . . . . .	50
3.5.	Reference classification uncertainty: geolocation and interpreter variability . . . . .	50
4.	Analysis . . . . .	51
4.1.	The error matrix . . . . .	51
4.2.	General principles of estimation for good practice . . . . .	51
4.3.	Estimating accuracy . . . . .	51
4.4.	Estimating area . . . . .	52
5.	Example of good practices: estimating area and assessing accuracy of forest change . . . . .	52
5.1.	Sampling design . . . . .	52
5.1.1.	Determining the sample size . . . . .	53
5.1.2.	Determine sample allocation to strata . . . . .	53
5.2.	Estimating accuracy, area and confidence intervals . . . . .	54
5.2.1.	Estimating accuracy . . . . .	54
5.2.2.	Estimating area and uncertainty . . . . .	54
6.	Summary . . . . .	54
6.1.	General . . . . .	54
6.2.	Sampling design . . . . .	55
6.3.	Response design . . . . .	55
6.4.	Analysis . . . . .	55
	Acknowledgments . . . . .	56
	References . . . . .	56

## 1. Introduction

Land change maps quantify a wide range of processes including wildfire (Schroeder, Wulder, Healey, & Moisen, 2011), forest harvest (Olofsson et al., 2011), forest disturbance (Huang et al., 2010), land use pressure (Drummond & Loveland, 2010) and urban expansion (Jeon, Olofsson, & Woodcock, 2013). Map users and producers are acutely interested in communicating and understanding the quality of these maps. Accordingly, guidance on how to assess accuracy of these maps in a consistent and transparent manner is a necessity. The use of remote sensing products depicting change for scientific, management, or policy support activities all require quantitative accuracy statements to buttress the confidence in the information generated and in any subsequent reporting or inferences made. Area estimation, whether of change in land cover/use or of status of land cover/use at a single date, is a natural value-added use of land change maps in many local, national and global land accounting applications. For example, the amount of land area allocated for a specific use is a key country reporting requirement to the United Nations (UN) Food and Agriculture Organization (FAO) statistics and the global forest resources' assessment (FAO, 2010) as well as for countries reporting under the Kyoto protocol and the evolving activities for the UN Collaborative Programme on Reducing Emissions from Deforestation and Forest Degradation – UN-REDD (Grassi, Monni, Federici, Achard, & Mollicone, 2008; UN-REDD, 2008). Estimates of forest extent or deforestation are often derived via remote sensing (cf. Achard et al., 2002; DeFries et al., 2002; Hansen, Stehman, & Potapov, 2010), and area estimation also plays a prominent role in ongoing efforts to establish scientifically valid protocols for forest change monitoring in the context of specific accounting applications to policy approaches for reducing greenhouse gas emissions from forests (DeFries et al., 2007; GOCF-GOLD, 2011).

A key strength of remote sensing is that it enables spatially exhaustive, wall-to-wall coverage of the area of interest. However, as might be expected with any mapping process, the results are rarely perfect. Placing spatially and categorically continuous conditions into discrete classes may result in confusion at the categorical transitions. Error can also result from the change mapping process, the data used, and analyst biases (Foody, 2010). Change detection and mapping approaches using remotely sensed data are increasingly robust, with improvements aimed at the mitigation of these sources of error. However, any map made from remotely sensed data can be assumed to contain some error, with the areas calculated from the map (e.g., pixel counting)

also potentially subject to bias. An accuracy assessment identifies the errors of the classification, and the sample data can be used for estimating both accuracy and area along with the uncertainty of these estimates. While the notion of accuracy assessment is well-established within the remote sensing community (Foody, 2002; Strahler et al., 2006), studies of land change routinely fail to assess the accuracy of the final change maps and few published studies of land change make full use of the information obtained from accuracy assessments (Olofsson, Foody, Stehman, & Woodcock, 2013).

### 1.1. Good practice recommendations

In this article, we synthesize the current status of key steps and methods that are needed to complete an accuracy assessment of a land change map and to estimate area of land change. This article addresses the fundamental protocols required to produce scientifically rigorous and transparent estimates of accuracy and area. The set of good practice recommendations provides guidelines to assist both scientists and practitioners in the design and implementation of accuracy assessment and area estimation methods applied to land change assessments using remote sensing. The accuracy and area estimation objectives are linked via a map of change. A change map provides a spatially explicit depiction of change and this spatial information can be readily aggregated to calculate the total mapped area or the proportion of mapped area of change for the region of interest (ROI). Accuracy assessment addresses questions related to how well locations of mapped change correspond to actual areas of change. A fundamental premise of the recommended good practices methodology is that the change map will be subject to an accuracy assessment based on a sample of higher quality change information (i.e., the reference classification). The higher quality reference classification is compared to the map classification on a location-specific basis to quantify accuracy of the change map and to estimate area. Although it is possible to estimate area of change without producing a change map (Achard et al., 2002; FAO, 2010; Hansen et al., 2010), we will assume that a map of change exists (although there will not necessarily be a map for each date). The focus for this document is change between two dates.

Before any detailed planning of the response and sampling designs is undertaken, a basic visual assessment should be conducted to identify obvious errors and concerns in the remotely sensed product. This assessment provides an evaluation of the map's suitability for the intended application and should detect if a map is so unsuitable for

use that there is no value in proceeding to a more detailed assessment. The visual assessment should also highlight errors that are easy to remove enabling the map to be refined prior to initiating a detailed assessment or confirm that no obvious concerns exist and the map is ready for further rigorous evaluation.

We separate the accuracy assessment methodology into three major components, the response design, sampling design, and analysis (Stehman & Czaplewski, 1998). The response design encompasses all aspects of the protocol that lead to determining whether the map and reference classifications are in agreement. Because it is often impractical to apply the response design to the entire ROI, a subset of the area is sampled. The sampling design is the protocol for selecting that subset of the ROI. The analysis includes protocols for defining how to quantify accuracy along with the formulas and inference framework for estimating accuracy and area and quantifying uncertainty of these estimates. A separate section of this guidance document is devoted to each of these three major components of accuracy assessment methodology. These sections are followed by an example of the recommended workflow.

### 1.2. Context of good practice recommendations

The good practice recommendations are intended to represent a synthesis of the current science of accuracy assessment and area estimation. We fully anticipate that improved methods will be developed over time. As the designation of “best practice” implies a singular approach, we prefer the use of “good practice” to indicate that “best” is relative and will vary, with one hard-coded approach not always appropriate. In communicating good practices, desirable features and selection criteria can be followed to ensure that the protocol applied satisfies – as thoroughly as possible – the accuracy and area estimation recommendations. The good practice recommendations do not preclude the existence of other acceptable practices, but instead represent protocols that, if implemented correctly, would ensure scientific credibility of the results. Furthermore, the recommendations presented herein allow flexibility to choose specific details of the different components of the methodology. For example, while the general recommendation for the sampling design is to implement a probability sampling protocol, there are numerous sampling designs that meet this criterion (Stehman, 2009). Similarly, the response design protocol allows flexibility to use a variety of different sources for determining the reference classification and multiple options exist for defining agreement between the map and reference classifications. The good practices recommendations represent an ideal to strive for, but it is likely that most projects will not satisfy every recommendation. Documenting and justifying deviations from good practices are expected features of many accuracy assessment and area estimation studies. For the most part, the good practice recommendations consist of methods for which there is considerable experience of practical use in the remote sensing community.

These good practice recommendations for area estimation and accuracy assessment of land change build on earlier guidelines for single-date land-cover maps described by Strahler et al. (2006). Strahler et al. (2006) presented general guiding principles of good practices with less emphasis on details of methodology. In the intervening years since Strahler et al. (2006), additional theory and practical application related to accuracy assessment and area estimation have been accumulated, and this current document avails upon these developments to delve more deeply into methodological details. We do not attempt to provide an exhaustive description of methods given the range of issues and the highly application-specific nature of the topic. Instead, our purpose is to focus upon the main issues needed to establish a common basis of good practice methodology that will be generally applicable and result in transparent methods and rigorous estimates of accuracy and area. A list of recommendations for all components of the process (sampling design, response design, and analysis) is presented in the Summary section (Section 6).

Estimating area and accuracy of change maps introduces additional methodological challenges that were not within the scope addressed by Strahler et al. (2006). In particular, the area estimation objective was not addressed at all by Strahler et al. (2006). Accuracy assessment of change highlights many unique challenges, including the dynamic nature of the reference data, and aspects of the change features including type, severity, persistence, and area. Another challenge is that change is usually a rare feature over a given landscape. The accuracy of a map and the area estimates derived with its aid are a function of the land-cover mosaic under study, the underlying imagery and the methods applied. Accuracy and area estimates for the same region will, for example, vary if using a per-pixel or object-based classification or if the spatial resolution of the imagery is altered (cf. Baker, Warner, Conley, & McNeil, 2013; Duro, Franklin, & Duba, 2012; Johnson, 2013).

Our recommendations also focus on methods for providing rigorous estimates of land (area) change and its uncertainties. A primary use of such estimates is in analysis and accounting frameworks such as national inventories. In evolving frameworks compensating for successful climate change mitigation actions in the forest sector (such as REDD +, DeFries et al., 2007), the consideration of uncertainties are likely linked with financial incentives and are subject to critical international political negotiations on reporting and verification (Sanz-Sanchez, Herold, & Penman, 2013). Understanding and management of uncertainties in area change is essential, particularly because data and capacity gaps in forest monitoring are large in many developing countries (Romijn, Herold, Kooistra, Murdiyarso, & Verchot, 2012). Accuracy assessments should also focus on identifying and addressing error sources, and prioritize on capacity development needs to provide continuous improvements and reduce uncertainties in the estimates over time. This also includes assessing the value of data streams from evolving monitoring technologies (de Sy et al., 2012; Pratihast, Herold, de Sy, Murdiyarso, & Skutsch, 2013) where the ultimate impact on lower uncertainties need to be proven in operational contexts. Thus, the methods of good practice presented here are generic for providing rigorous estimates, and having agreed-upon tools to do so will provide the saliency and legitimacy for using them in quantifying improvements in monitoring systems, and for dealing with uncertainties in financial compensation schemes (e.g., for climate change mitigation actions).

This article synthesizes key steps and methods needed to complete an accuracy assessment of a change map and to estimate area and accuracy of the map classes. It addresses the protocols required to produce scientifically rigorous and transparent estimates of accuracy and area.

## 2. Sampling design

The sampling design is the protocol for selecting the subset of spatial units (e.g., pixels or polygons) that will form the basis of the accuracy assessment. Choosing a sampling design requires a consideration of the specific objectives of the accuracy assessment and a prioritized list of desirable design criteria. The most critical recommendation is that the sampling design should be a probability sampling design. An essential element of probability sampling is that randomization is incorporated in the sample selection protocol. Probability sampling is defined in terms of inclusion probabilities, where an inclusion probability relates the likelihood of a given unit being included in the sample (Stehman, 2000). The two conditions defining a probability sample are that the inclusion probability must be known for each unit selected in the sample and the inclusion probability must be greater than zero for all units in the ROI (Stehman, 2001).

A variety of probability sampling designs are applicable to accuracy assessment and area estimation, with the most commonly used designs being simple random, stratified random, and systematic (Stehman, 2009). Non-probability sampling protocols include purposely selecting sample units (e.g., choosing units that are convenient to access), restricting the sample to homogeneous areas, and implementing a complex or ad hoc selection protocol for which it is not possible to derive the

inclusion probabilities. The condition that the inclusion probabilities must be known for the units selected in the sample must be adhered to. These inclusion probabilities are the basis of the estimates of accuracy and area, so if they are not known, the probabilistic basis for design-based inference (see Section 4.2) is forfeited. It is difficult to envision a circumstance in which a deviation from this condition of probability sampling (i.e., known inclusion probabilities) would be acceptable for a scientifically rigorous assessment of accuracy.

In practice, it is not always possible to adhere perfectly to a probability sampling protocol (Stehman, 2001). For example, if the response design specifies field visits to sample locations, it may be too dangerous or too expensive to access some of the sample units. Conversely, persistent cloud coverage or lack of useable imagery for portions of the ROI may prevent obtaining the reference classification for some sample units. The reference data are often derived from another set of imagery and the spatial and temporal coverage of reference data might be different from the coverage of the imagery used to create the map. If the reference classification for a sample unit cannot be obtained, the inclusion probability is zero for that unit. All deviations from the probability sampling protocol should be documented and quantified to the greatest extent possible. For example, the proportion of the selected sample units for which cloud cover prevented assessment of the unit should be reported, or the proportion of area of the ROI for which the reference imagery is not available should be documented. Whereas probability sampling ensures representation of the population via the rigorous probabilistic basis of inference established, when a large proportion of the ROI is not available to be sampled, the question of how well the sample represents the population must be addressed by subjective judgment.

### 2.1. Choosing the sampling design

The major decisions in choosing a sampling design relate to trade-offs among different designs in terms of advantages to meet specified accuracy objectives and priority desirable design criteria. The objectives commonly specified are to estimate overall accuracy, user's accuracy (or commission error), producer's accuracy (or omission error), and area of each class (e.g., area of each type of land change). Estimates for subregions of the ROI are also often of interest (cf. Scepán, 1999). Desirable sampling design criteria include: probability sampling design, ease and practicality of implementation, cost effectiveness, representative spatial distribution across the ROI, small standard errors in the accuracy and area estimates, ease of accommodating a change in any step in the implementation of the design, and availability of an approximately unbiased estimator of variance. Determining whether any or all of these desirable design criteria have been satisfied by the chosen sampling design may be subjective. For example, determining what constitutes a small standard error will depend on the application and may vary for different estimates within the same project. There are also precedents for defining an accuracy target and desired error bounds as a means for determination of sample size using standard statistical theory (Wulder, Franklin, White, Linke, & Magnussen, 2006) (see also Section 5.1.1).

Stehman and Foody (2009) provide an overview and comparison of the basic sampling designs typically applied to accuracy assessment. Stehman (2009) provides a more expansive review of sampling design options and discusses how these designs fulfill different objectives and desirable design criteria. A variety of sampling designs will satisfy good practice guidelines so the key is to choose a design well suited for a given application. Three key decisions that strongly influence the choice of sampling design are whether to use strata, whether to use clusters, and whether to implement a systematic or simple random selection protocol (Stehman, 2009). Each of these decisions will be discussed in the following subsections.

#### 2.1.1. Strata

There is often a desire to partition the ROI into discrete, mutually exclusive subsets or strata (e.g., a global map could be stratified

geographically by continents). Stratification is a partitioning of the ROI in which each assessment unit is assigned to a single stratum. The two most common attributes used to construct strata are the classes determined from the map and geographic subregions within the ROI. Stratification is implemented for two primary purposes. The first purpose is when the strata are of interest for reporting results (e.g., accuracy and area are reported by land-cover class or by geographic subregion). The second use of stratification is to improve the precision of the accuracy and area estimates. For example, when strata are created for the objective of reporting accuracy by strata, the stratified design allows specifying a sample size for each stratum to ensure that a precise estimate is obtained for each stratum. Land change often occupies a small proportion of the landscape, so a change stratum can be identified and the sample size allocated to this stratum can be large enough to produce a small standard error for the change user's accuracy estimate.

The practical reality is that limited resources will likely be available for the reference sample and this constraint will strongly impact sample allocation decisions because different allocations favor different estimation objectives. For example, allocating equal sample sizes to all strata favors estimation of user's accuracy over estimation of overall and producer's accuracies (Stehman, 2012). Conversely, the standard errors for estimating producer's and overall accuracies are typically smaller for proportional allocation (i.e., the sample size allocated to each stratum is proportional to the area of the stratum) relative to equal allocation. As a compromise between favoring user's versus producer's and overall accuracies, the allocation recommended is to shift the allocation slightly away from proportional allocation by increasing the sample size in the rarer classes, but the sample size for the rare classes should not be increased to the point where the final allocation is equal to allocation (see Section 5 for an example). The sample size allocation decision can be informed by calculating the anticipated standard errors (see Sections 4.3 and 4.4) for different sample sizes and different allocations. An ineffective allocation of sample size to strata will not result in biased estimators of accuracy or area, but it may result in larger standard errors (see Section 5 for an example).

When stratified sampling is applied to a single date land-cover map, it is usually feasible to define a stratum for each land-cover class (Wulder, White, Magnussen, & McDonald, 2007). Identifying an effective stratification for change can be more challenging. A common approach is to use a map of change to identify the strata, and such strata are effective for estimating user's accuracy of change precisely. However, the number of different types of change may be so large that defining every change type as a stratum is not advisable. For example, in a post-classification comparison of two land-cover maps that each include 8 land-cover classes, there are 56 possible types of change in the final change map. If each stratum must receive a relatively large sample to achieve a precise user's accuracy estimate, the overall sample size may be unaffordable.

The trade-offs between precision of user's accuracy, producer's accuracy, and area estimates from different sample size allocations become exacerbated as the number of strata increases. Some types of change may be very unlikely to occur and consequently could be eliminated as strata. To further reduce the number of strata, strata could be defined on the basis of generalized change categories (Wickham et al., 2013). For example, a stratum could be change from any class to urban (i.e., urban gain), and another stratum could be change to any class from forest (i.e., forest loss). These generalized or aggregated change strata are obviously less focused on all possible individual change types. For example, the forest loss stratum could include forest to developed, forest to water, or forest to cropland. These generalized change strata would allow for specifying the sample size allocated to different general change types, but within one of the generalized strata, the sample size allocated to the individual change types would be proportional to the area of that change type. For example, if the most common type of forest loss is to cropland and the least common change is forest loss to water, many more of the sample units within the forest loss stratum will be forest-to-cropland-conversion. Strahler et al. (2006, Fig. 5.2, p. 32) provides

additional examples of aggregated change classes that could be used as strata.

The desire to limit the number of strata motivates discussion of subpopulation estimation as it relates to sampling design. A subpopulation is any subset of the ROI, for example a particular type of change or a particular subregion. Subpopulations can be defined as strata, but it is not necessary for a subpopulation to be defined as a stratum to produce an estimate for that subpopulation. For example, when aggregating multiple types of change into a generalized change stratum, it would still be possible to estimate accuracy of each of the subpopulations representing the individual types of change making up the aggregated change stratum. However, if these subpopulations are not defined as strata, the sample size representing the subpopulation may not be large enough to obtain a precise estimate. Resources available for accuracy assessment may require limiting the number of strata used in the design, so prioritizing subpopulations may be necessary to establish which subpopulations are defined as strata.

It is sometimes the case that several maps will be assessed based on a common accuracy assessment sample. This forces a decision on whether the strata should be based on a single map (and if so, which map) or if the strata should be defined by a combination of the multiple maps. Once strata are defined and the sample is selected using these strata, the strata become a fixed feature of the design because the analysis is dependent on the estimation weight associated with each sample unit and this weight is determined by the sampling design. Fortunately, whatever the decision is to define strata when multiple maps are to be assessed, the sample reference data are still valid to assess any of the maps, even if the strata are defined on the basis of a single map. The principles of estimation outlined in the Analysis section (Section 4) must be adhered to, and this simply requires using the estimation weights for the sample units determined by the original stratified selection protocol. The impact of the choice of strata will be reflected in the standard errors of the estimates. Olofsson et al. (2012) and Stehman, Olofsson, Woodcock, Herold, and Friedl (2012) discuss sampling design issues associated with constructing a reference validation database that would allow assessment of multiple maps.

To summarize the recommendations related to the important question of whether to incorporate stratification in the sampling design, stratifying by mapped change and by subregions is justified to achieve the objective of precise class-specific accuracy and to report accuracy by subregion. If the overall sample size is not adequate to support both class-specific and subregion accuracy estimates, the subregional stratification may be omitted and accuracy by subregion relegated to the status of subpopulation estimation. The recommended allocation of sample size to the strata defined by the map classes is to increase the sample size for the rarer classes making the sample size per stratum more equitable than what would result from proportional allocation, but not pushing to the point of equal allocation. The rationale for this recommendation is that user's accuracy is often a priority objective and we can control the precision of the user's accuracy estimates by the choice of sample allocation. However, the trade-off is that a design allocation chosen solely for the objective of user's accuracy precision (i.e., equal allocation) may be detrimental to precision of estimates of overall accuracy, producer's accuracy, and area, so a compromise allocation is in order. Lastly, defining aggregations of change types as strata may be necessary if the number of strata needs to be limited, and accuracy and area estimates for the individual change types would be obtained as subpopulation estimates.

### 2.1.2. Cluster sampling

A cluster is a sampling unit that consists of one or more of the basic assessment units specified by the response design. For example, a cluster could be a  $3 \times 3$  block of 9 pixels or a  $1 \text{ km} \times 1 \text{ km}$  cluster containing 100 1 ha assessment units. In cluster sampling, a sample of clusters is selected and the spatial units within each cluster are therefore selected as a group rather than selected as individual entities. Each of the spatial

units within a cluster is still interpreted as a separate unit even though it is selected into the sample as part of a cluster. For example, a  $3 \times 3$  pixel cluster would require obtaining the reference classification for individual pixels within the cluster.

The primary motivation for cluster sampling is to reduce the cost of data collection. For example, if field visits are required to obtain the reference classification, transit time and costs may be reduced if the sample units are grouped spatially into clusters. Zimmerman et al. (2013) used cluster sampling to reduce the number of raster images (i.e., clusters) required because the primary cost of the sampling protocol was associated with processing the very high resolution images used for reference data. As another example, Stehman and Selkowitz (2010) used a  $27 \text{ km} \times 27 \text{ km}$  cluster sampling unit to constrain sample locations to a single day of flight time per cluster when the reference data were collected by aircraft. Cluster sampling may also be motivated by the objectives of an accuracy assessment. For example, a cluster sampling unit becomes necessary to assess accuracy at multiple spatial supports (e.g., single pixel, 1 ha unit, and  $1 \text{ km}^2$  unit).

The cost savings gained by cluster sampling should be substantial before choosing this design because the correlation among units within a cluster (i.e., intraclass correlation) often reduces precision relative to a simple random sample of equal size. Focusing on the specific example of estimating land-cover area in Europe, Gallego (2012) showed that a  $10 \text{ km} \times 10 \text{ km}$  sampling unit produced equivalent information to that of a simple random sample of only 25 points or fewer. The low yield of information per cluster diminishes the cost advantage of cluster sampling if the intraclass correlation is high. Another potential disadvantage of cluster sampling is that it complicates stratification when the strata are the map classes and the assessment unit is a pixel. In the simplest setting, each cluster would be assigned to a stratum, but rules have to be established for assigning a cluster to a stratum when the cluster includes area of several different classes. Cluster sampling can be combined with stratification of pixels by the map class of each pixel in a two-stage stratified cluster sampling approach (Stehman, Sohl & Loveland, 2003; Stehman, Wickham, Wade & Smith, 2008), but such designs require more complex analysis and implementation protocols than what are required of a stratified design without clusters. Because of the added complexity cluster sampling introduces for sampling design (e.g., accommodating stratification within a cluster sampling design) and estimation (e.g., estimating standard errors), we recommend this design only in cases for which the objectives require a cluster sampling unit or in which the cost savings or practical advantages of cluster sampling are substantial.

### 2.1.3. Systematic vs. random selection

The two most common selection protocols implemented in accuracy assessment are simple random and systematic sampling (we define "systematic" as selecting a starting point at random with equal probability and then sampling with a fixed distance between sample locations). Both protocols can be implemented to select units from within strata or to select clusters, and both can be applied to a ROI that is not partitioned into strata or clusters. Unbiased estimators of the various accuracy parameters are available from either systematic or simple random selection, so the bias criterion is not a basis for choosing between these options. Instead, the choice of simple random versus systematic depends on how each selection protocol satisfies the priority desirable design criteria (Stehman, 2009). For example, systematic sampling is often simpler to implement when the response design is based on field visits, but the greater convenience of systematic versus simple random is diminished when working with imagery or aerial photographs as a source of the reference data. Typically, systematic selection will yield more precise estimates than simple random selection, but systematic sampling requires use of a variance approximation so if unbiased variance estimation is a priority criterion, simple random is preferred. Simple random selection also is advantageous if it is likely that the sample size will need to be modified during the course of the accuracy

assessment (Stehman et al., 2012). A scenario in which systematic selection opportunistically arises is when accuracy assessment reference data can be simultaneously obtained in conjunction with another field sampling activity. For example, many national forest inventories employ a systematic sample of field plots (Tomppo, Gschwantner, Lawrence, & McRoberts, 2010) and these field plot data may be an inexpensive, high quality source of reference data. In general, the simple random selection protocol will better satisfy the desirable design criteria and is the recommended option. However, systematic selection is also nearly always acceptable.

## 2.2. A recommended good practice sampling design

Stratified random sampling is a practical design that satisfies the basic accuracy assessment objectives and most of the desirable design criteria. Stratified random sampling affords the option to increase the sample size in classes that occupy a small proportion of area to reduce the standard errors of the class-specific accuracy estimates for these rare classes. Thus this design addresses the key objective of estimating class-specific accuracy. In regard to the desirable design criteria, stratified random sampling is a probability sampling design and it is one of the easier designs to implement. Stratified sampling is commonly used in accuracy assessment so it has an advantage of being familiar to the remote sensing community (cf. Cakir, Khorram, & Nelson, 2006; Huang et al., 2010; Mayaux et al., 2006; Olofsson et al., 2011). Increasing or decreasing the sample size after the data collection has begun is readily accommodated by stratified random sampling, and unbiased variance estimators are available thus avoiding the need to use variance approximations. An assumption implicit in this recommendation is that change between two dates is of interest. Little work has been done to investigate the effective use of strata for multiple change periods. In the case of stratification based on a change map, it is assumed that reference data for the sampled locations exists for the initial date of the change period (e.g., archived imagery or aerial photography is available). If the reference data must be obtained in real time (e.g., via ground visit), it would not be possible to stratify by a change map that does not yet exist at the initial date. An alternative would be to stratify by anticipated change or predicted change, with the effectiveness of such strata dependent on how well the predicted change matched with the ensuing reality of change.

## 3. Response design

For the accuracy assessment objective, the response design encompasses all steps of the protocol that lead to a decision regarding agreement of the reference and map classifications. For area estimation, the response design provides the best available classification of change for each spatial unit sampled. The four major features of the response design are the spatial unit, the source or sources of information used to determine the reference classification, the labeling protocol for the reference classification, and a definition of agreement. Each of these major features is discussed in the following subsections.

### 3.1. Spatial assessment unit

The spatial unit that serves as the basis for the location-specific comparison of the reference classification and map classification can be a pixel, polygon (or segment), or block (Stehman & Wickham, 2011). The ROI is partitioned based on the chosen spatial unit (i.e., the region is completely tiled by these non-overlapping spatial units). Commonly, the pixel is selected as the spatial unit. The pixel is an arbitrary unit defined mainly by the properties of the sensing system used to acquire the remotely sensed data or a function of the grid used to sub-divide space in a raster based data set. A polygon is defined as a unit of area, perhaps irregular in shape, representing a meaningful feature of land cover. For example, a polygon may be delineated from a map such that

the area within the polygon has the same map classification (e.g., the entire polygon is stable forest or the entire polygon represents an area of change from forest to urban). Polygons defined on the basis of a map will be called “map polygons.” Alternatively, a polygon could be delineated on the basis of the reference classification as an area within which the reference class is the same. A polygon delineated on the basis of the reference classification will be called a “reference polygon”. A “block” spatial assessment unit is defined as a rectangular array of pixels (e.g., a  $3 \times 3$  block of pixels). Irrespective of the spatial unit selected, it is important to note that some spatial units may be impure, i.e., they represent an area of more than one class. Mixed pixels are common, especially in coarse spatial resolution data. Similarly, it is possible that a map polygon is not internally homogeneous in terms of the reference classification, and a reference polygon may not be internally homogeneous in terms of the map classification. A polygon defined by a segmentation algorithm would not necessarily be homogeneous in terms of either the map or the reference classifications.

Pixels, polygons, or blocks can be used as the spatial unit in accuracy assessment. Regardless of the unit chosen, a critical feature of the response design protocol is that the spatially explicit character of the accuracy assessment must be retained. Practitioners should aim to have reference data with an equal or finer level of detail than the data used to create the map, but we make no recommendation regarding the choice of spatial assessment unit. However, once the spatial assessment unit has been chosen, there will be good practice recommendations associated with that specific unit and the choice of spatial unit also has implications on the sampling design (Stehman & Wickham, 2011) and analysis. Estimates of accuracy and area derived from the same map but through the use of different spatial units may be unequal.

### 3.2. Sources of reference data

The reference classification can be determined from a variety of sources ranging from actual ground visits to the sample locations or the use of aerial photography or satellite imagery. There are two ways to ensure that the reference classification is of higher quality than the map classification: 1) the reference source has to be of higher quality than what was used to create the map classification, and 2) if using the same source material for both the map and reference classifications, the process to create the reference classification has to be more accurate than the process used to create the classification being evaluated. For example, if Landsat imagery is used to create the map and Landsat is the only available imagery for the accuracy assessment, then the process for obtaining the reference classification has to be more accurate than the process for obtaining the map classification. Additionally, other spatial data may be used to improve the quality of the reference classification, such as forest inventory data or some form of vector data (e.g., roads, pipelines, or crop records). In this subsection, different potential sources of reference data for assessing accuracy of change are identified and strengths and weaknesses of these sources are described.

Possible reference data sources include field plots, aerial photography, forest inventory data, airborne video, lidar, and satellite imagery (Table 1). Additional sources of freely accessible reference data may also be opportunistically available from data mining and crowdsourcing (Foody & Boyd, 2013; Iwao, Nishida, Kinoshita, & Yamagata, 2006).

**Table 1**  
Possible reference data sources.

Reference data source	Exemplar citation
Field plots	Hyypä et al. (2000)
Air photography	Skirvin et al. (2004)
Forest inventory data	McRoberts (2011); Wulder, White, et al. (2006)
Airborne video	Wulder et al. (2007)
Lidar	Lindberg, Olofsson, Holmgren, and Olsson (2012)
Satellite imagery	Scepan (1999); Cohen et al. (2010)
Crowdsourcing	Iwao et al. (2006); Foody and Boyd (2013)



Practical considerations regarding costs often influence the selection of reference data, or the use of existing data. While existing or lower cost data may be desirable from a purchase perspective, the use of disparate data sources will result in additional effort by project analysts to deal with exceptions and inconsistencies. A key to using disparate data sources is to have the reference data that are actually used in the accuracy assessment be, as much as possible, invariant to source. For example, the creation of attributed change polygons makes the polygon the common denominator, rather than the source data. Creating polygonal change units in a portable format and populating a minimum set of fields to support a consistent labeling protocol is desirable. The information to be recorded for each change unit is itemized in Table 2.

Ideally a data source is available for the entire ROI, representing the change types and dates of interest, at a low cost. The realities versus the ideal result in a series of considerations that are detailed in Table 3. For instance, if the ROI is small, cost may be less of an issue and access may not be relevant. For large area projects over poorly monitored areas, existing data sources are not often available so data purchase and interpretation costs become the dominant criteria. The ease of interpretation and consistency of source reference data permits economies in the project flow for the analysts and also promotes automation of repeated activities. Further, the development of a well-documented and consistent change validation data set will have utility for multiple projects and purposes.

Both high- and very high spatial resolution satellite data are viable candidates for reference data. Imagery is typically considered as very high spatial resolution (VHSR) with a spatial resolution of <1 m and high spatial resolution (HSR) with a spatial resolution of <10 m. Both data sources provide information that is finer than the data used in most large area monitoring projects, which would typically have a spatial resolution of greater than 10 m. At the fine spatial resolution of satellite-borne VHSR imagery, panchromatic is often the only spectral information collected. The typical 400 to 900 nm panchromatic data with small pixels (0.50 m in the case of WorldView-1) closely resemble large scale aerial photography and can be interpreted using established aerial photograph interpretation techniques (Wulder, White, Hay, & Castilla, 2008) or subject to digital analyses (cf. Falkowski, Wulder, White, & Gillis, 2009). Both the SPOT Image® and DigitalGlobe® archives can be accessed through Google Earth™, with the image extents by year portrayed. The presence of freely accessible high spatial resolution imagery online through Google Earth™ also presents low cost interpretation options. Limitations of this approach include a lack of data prior to the initiation of the high spatial resolution satellite commercial era (circa 2000), spatial distribution of available imagery, and the actual temporal revisit of the images available. The reported temporal revisit can be on the order of days based upon an ability to point the sensor head. For instance, IKONOS has off-nadir revisit of 3 to 5 days, with 144 days

required for nadir revisit (Wulder, White, Coops, & Butson, 2008). The implication is that when the sun-surface-sensor viewing geometry changes the structure captured changes, such that trees evident on one image may be occluded in another. For a given on-line accessible source of satellite imagery, it should not be expected that historical, archival, global coverage from launch to present exist. Regardless, the ability to view images from multiple years can help determine that date when a change (e.g., a disturbance) occurred. The additional context provided around particular change events aids with interpretation of change type (e.g., determination of harvesting versus forest removal in support of agricultural expansion).

There are few, if any, reference data sources that are available with a uniform likelihood globally. There are some archival datasets with wide global coverage (e.g., Kompsat); although, the utility of these data sets may be limited. The utility of any given reference data source when used to capture and relate change is the date or represented by the data. While less of an issue with satellite data, air photos and maps may not be of a known vintage. Acquisition dates of historic photos are often lost, plus maps are often representative of a period, not a singular date. Knowing the conditions that previously existed may not be helpful if the date of change occurrence is not known.

Over some regions, land use change and silvicultural records may also be available to inform on the land-cover change. Note that forest harvesting is a land-cover change relating a successional stage, rather than a land use change (which implies a permanent change in how a particular parcel of land is used — e.g., forestry to agriculture). This distinction is important for both monitoring and reporting purposes as the permanent removal of forests has differing carbon consequences than forest harvesting (Kurz, 2010).

While the good practice guidelines advocate for use of reference data of finer spatial resolution than the map product, this is especially so for single date interpretations of the reference data. Following the opening of the Landsat archive by the USGS (Woodcock et al., 2008), time series of imagery created new opportunities for using imagery of the same spatial resolution (e.g., Landsat) when archival data are available. Simple visual approaches may be applied, such as in Fig. 1, where a change event (fire) that is evident in 2010 can be timed quite precisely by the evidence captured (smoke plume) showing when the fire occurred. This type of change dating is rather opportunistic and not to be commonly expected.

A more reliable means for determining the timing of change events can be from developing and interrogating time series of images (Kennedy, Yang, & Cohen, 2010). To ensure the quality of time series transitions developed, Cohen, Yang, and Kennedy (2010) created a logic and tool for determining the timing and nature of changes captured (TimeSync, <http://timesync.forestry.oregonstate.edu/>). Based upon the image collection and archiving protocols present through the history of Landsat, the spatial and temporal coverage of imagery is not uniform. The temporal precision possible for dating changes based upon time series analysis is likely weaker for locations that already have a paucity of data. This situation is due to the historic practices followed at given Landsat receiving stations through to the commercial era (during the 1980s) when fewer images were collected and archived (Wulder, Masek, Cohen, Loveland, & Woodcock, 2012). It should not be assumed that the temporal density possible for the conterminous United States is possible for all other regions (Schroeder et al., 2011).

Another critical aspect of the response design is that the change period represented by the reference classification must be synchronous with the change period of the classification. Consider a map representing change between 2000 and 2010. To capture the northern hemisphere peak photosynthetic period, the imagery used for this hypothetical project was collected July 15, 2000, and 10 years later, July 15, 2010. The reference data should be collected in 2010, but ideally not after July 15 (assuming similar satellite overpass times) to avoid confusion. Data collected after July 15, 2010 will have to be vetted to ensure that the change present in the reference data did not occur

**Table 2**

Example characteristics to record for each change polygon. Some attributes can be generated in the GIS; others will need to be entered by the analyst. Notion is that information is captured and carried to provide insights and a record regarding the changes captured. The aim is that the change polygons can be used in a manner that is invariant to source, but that metadata is captured to explain or better understand any data related anomalies that may emerge.

Attribute	Definition/comments
Change area	Area changed, e.g., polygon size in hectares
Change perimeter	Perimeter of polygon, in meters
Change type	Notation of change type, harvest, fire, insect, urban expansion, agricultural development
Change date	As possible, note the change date. May be available from other records, e.g., when a fire occurred, or the acquisition date of the image or photography used.
Data source	Note the data source from which the change polygon is made
Analyst	Name or code to denote the interpreter
Date interpreted	Note the date when the interpretation occurred

**Table 3**  
Elements for consideration when selecting reference data.

Element	Considerations
Cost	What is the budget? What amount per unit of reference data can be purchased? Is the interpretation/labeling protocol efficient?
Ease of access	Varies by data type. Can field visits be made? Is archival image data available?
Ease of use	Is the data produced in a consistent fashion? Is it in formats that are commonly used?
Opportunity for consistency	Can protocols be developed and applied in a systematic and repetitive fashion? Can some tasks be automated?
Vintage – temporal representation	Is the data representative of a time or time period that is relevant to the change product under consideration?
Spatial coverage	Are there opportunities for multiple reference sites from a given reference data source?
Interpretability of change types	Does the data source capture and portray the change types of interest? E.g., is the spatial resolution sufficiently fine to enable interpretation?
Geolocation	Can the candidate reference data source be assumed to be accurately positioned? Will additional geolocation activities be required?

after the product date of the change map. Imagery from the same year is desired but may not always be possible. As such, it is required that the change reference data approximates the date the change occurred as precisely as available. Multiple images help refine the timing of the change event. Mismatched change periods between the map and reference classifications would be a major source of reference data error.

### 3.3. Reference labeling protocol

The labeling protocol refers to the steps in the response design that take the information provided by the reference data and convert that information to the label or labels constituting the reference classification. Labeling is far from trivial with numerous definitions for land-cover classes in use (cf. Comber, Wadsworth, & Fisher, 2008) although recent developments such as the FAO's Land Cover Classification system (LCCS) may act to enhance interoperability (Ahlqvist, 2008). The labeling protocol should also include specification of a minimum mapping unit (MMU) for the reference classification. The MMU can have important implications for accuracy assessment and area estimation. For example, increasing the size of the MMU will lead to a reduction in the representation of classes that occupy small, often fragmented, patches (Saura, 2002). Changing the MMU can also impact accuracy estimates, although the effect is most apparent when a large change is made (Knight & Lunetta, 2003). Small patches present a challenge to mapping (cf. He, Franklin, Guo, & Stenhouse, 2011) and the accuracy of their mapping will degrade as the MMU is increased. However, it is possible that overall map accuracy may increase with a larger MMU, making it important to ensure that attention is focused on an appropriate measure of accuracy for the application in-hand. The precise effects of the MMU will vary as a function of the land-cover mosaic under study and the imagery used. The MMU specified for the response design does not necessarily have to match the MMU specified for the map. In fact, if the reference classification is intended to apply to a variety of maps, it would be likely that the MMU of the reference classification does not

match the map classification for all maps that might be assessed. Often the reference imagery or information will permit distinguishing smaller patches or features that can be distinguished from the map so a smaller MMU will be possible for the reference classification.

The easiest case for the labeling protocol occurs when the assessment unit is homogeneous and a single reference class label can be assigned (the reference class could be a type of change). Often, however, the situation will be more complex making class labeling less certain. For example, the assessment unit may contain a mixture of classes, and even if the unit is homogeneous, it may be difficult to assign a single label (e.g., change type) because the unit is not unambiguously one of the classes in the legend but instead falls between two of the discrete class options in the legend (i.e., land-cover classes are a continuum represented on a discrete scale). A variety of options exist for labeling a unit when a single reference label does not adequately represent the uncertainty of a unit. One or more alternate reference class labels can be assigned to account for ambiguity in the reference classification. Another option when defining agreement is to construct a weighted agreement based on how closely the different classes are related. For example, in the GlobCover assessment, a “matrix” of class relationships was established (Mayaux et al., 2006, GLC2000). A fuzzy reference labeling protocol may also be employed, such as the linguistic scale devised by Gopal and Woodcock (1994) or a fuzzy membership vector in which the reference label for a unit specifies a membership value for each class (Binaghi, Brivio, Ghezzi, & Rampini, 1999; Foody, 1996). Another option for mixed units is to specify the proportion of area of each class present in the unit (Foody, Campbell, Trodd, & Wood, 1992; Lewis & Brown, 2001). A different characterization of uncertainty in the reference classification is obtained by assigning a confidence rating that represents the interpreter's perception of uncertainty in the reference classification for that unit. For example, low, moderate and high confidence ratings would indicate increasing confidence on the part of the interpreter that the reference classification is correct. Typically this information can then be used in the analysis to subset results by



**Fig. 1.** Landsat data can be used for the visual dating of change, with the fire event in progress in inset A, August 3, 2010, with the burned forest outcome evident in inset B, September 20, 2010, Yukon, Canada (Landsat Path 55, Row 18).

confidence rating (Powell et al., 2004; Wickham, Stehman, Fry, Smith, & Homer, 2001, Table 4).

The response design should include protocols to enhance consistency of the reference class labeling. For example, interpretation keys should be created if visual assessment is used to obtain the reference classification (Kelly, Estes, & Knight, 1999) and specific instructions to translate quantitative field data into reference labels should be provided and documented. If multiple interpreters are used, training interpreters to ensure consistency is critical. Interpreters should be in communication throughout the process to discuss and review difficult cases and to agree upon a common approach to labeling such cases. Difficult cases should be noted for future reference and consensus development (e.g., the imagery is retained and accessible, and the decision process leading to the reference label of the case is documented). Rather than solely visual approaches, entire high spatial resolution images can be classified, with the underlying imagery also maintained and accessible as support information to the accuracy assessment (that is, to gain/ensure confidence in the categories selected for a given location).

### 3.4. Defining agreement

Once the map and reference classifications have been obtained for a given spatial unit, rules for defining agreement must be specified before proceeding to the analyses that quantify accuracy. In the simplest case, a single class label is present for the map and a single label is provided by the reference classification. If these labels agree, the map class is correct for that unit; if the labels disagree, the type of misclassification is readily identified. Defining agreement becomes more complex if the assessment unit is not homogeneous or if more than a one class label is assigned by the map or reference classification. For example, if the reference classification provides a primary and secondary reference label, agreement can be defined as a match between the map label and either the primary or secondary reference label. If the reference classification consists of a vector of proportions of area of the classes present in the assessment unit (e.g., the area proportions of the classes are 0.2, 0.5, and 0.3), agreement can be defined as the proportion of area for which the map and reference labels are the same. The critical feature of the protocol for defining agreement is that it allows construction of an error matrix in which the elements of the matrix represent proportion of area of agreement and disagreement between the map and reference classifications. These proportions (in terms of area) achieve the necessary spatially explicit assessment of map accuracy and the requirements for area estimation.

### 3.5. Reference classification uncertainty: geolocation and interpreter variability

In an ideal case, the reference classification is based on a reference data set of such quality that the sample labels represent the ground truth (i.e. a “gold standard” reference data set). However, the reference classification is subject to uncertainty, and an assessment of this uncertainty should be conducted. Small errors in the reference data set can lead to large biases of the estimators of both classification accuracy and class area (Foody, 2010, 2013). Two potential sources of uncertainty in

the reference classification are the uncertainty associated with spatial co-registration of the map and reference location (Pontius, 2000) and uncertainty associated with the interpretation of the reference data (Pontius & Lippitt, 2006).

Geolocation error is defined as a mismatch between the location of the spatial assessment unit identified from the map and the location identified from the reference data. The response design should be constructed to minimize geolocation error. For instance, it is common for plots to have a GPS position. The quality of the GPS position can be related to the type of instrument used, which can provide an indication of spatial precision. The length of time, number of position measures to resolve the location, and the number of satellites are also aspects that can be recorded. The magnitude of geolocation error should be characterized by documenting the spatial location quality of the map and reference data sources (e.g., GPS units, aerial photography, or satellite imagery). If airborne imagery is to be used, aircraft positioning and pointing information should be collected. The GPS location of the aircraft does not necessarily indicate the position of the point on the ground that is captured in photographic or video data. A slight roll of the aircraft can create a mismatch between the recorded and actual positions. Error in the classification may be incorrectly indicated due to these spatial mismatches, especially for smaller change events or rare classes.

Interpreter uncertainty can be separated into two parts: 1) interpreter bias is defined as an error in the assignment of the reference class to the spatial unit; 2) interpreter variability is a difference between the reference class assigned to the same spatial unit by different interpreters (i.e., interpreter variability is the complement of among interpreter agreement). Ideally an assessment of both interpreter bias and interpreter variability would be conducted; in practice, assessing only interpreter variability may be feasible. The difficulty hindering assessment of interpreter bias is whether a “gold standard” of truth exists against which the interpreted reference classification can be compared. For example, on-the-ground reference data may serve to establish the gold standard of truth for land cover at a single date, but a gold standard for change based on field visits would be much more difficult and costly to establish. Comparison of interpreters to an “expert” interpreter is a practical but less satisfying option for quantifying interpreter bias and the success of this approach depends on how closely the expert classification mimics the gold standard. A distinction between the accuracy assessment of land cover and change does exist, whereby the continuous nature of land cover benefits more from field visits. Depending on the change categories of interest, field visits may not be as informative. For example, slower continuous changes may benefit from field visits, but rapid stand replacing disturbances may not. The date of change, if not captured in silvicultural records or fire maps, may actually be better captured from imagery of known vintage than through field visits (Cohen et al., 2010).

A number of issues arise when using multiple interpreters to obtain the reference classification (Wulder et al., 2007). Disagreements among interpreters evaluating the same sampling unit are likely. These disagreements may be resolved by a consensus agreement on the reference class; for example, Powell et al. (2004) required five interpreters to agree upon a specific class, with the outcome then treated as a “gold standard”. Constant communication among the multiple interpreters to discuss and document difficult cases is important to foster enhanced consistency and accuracy of the reference labeling process (Wickham et al., 2013).

The response design protocols described in this section have focused on land-cover changes that can be characterized by a complete change in class type: conversions of cover. In some studies attention is focused on more subtle changes or modifications of land cover, as changes in land cover can be considered as processes (Gómez, White, & Wulder, 2011) with gains and losses in vegetation captured and possible to assign a label (Kennedy et al., 2010). Cohen et al. (2010) show how investigation of time series of satellite imagery supported by period

**Table 4**  
Population error matrix of four classes with cell entries ( $p_{ij}$ ) expressed in terms of proportion of area as suggested by good practice recommendations.

		Reference				Total
		Class 1	Class 2	Class 3	Class 4	
Map	Class 1	$p_{11}$	$p_{12}$	$p_{13}$	$p_{14}$	$p_{1\cdot}$
	Class 2	$p_{21}$	$p_{22}$	$p_{23}$	$p_{24}$	$p_{2\cdot}$
	Class 3	$p_{31}$	$p_{32}$	$p_{33}$	$p_{34}$	$p_{3\cdot}$
	Class 4	$p_{41}$	$p_{42}$	$p_{43}$	$p_{44}$	$p_{4\cdot}$
	Total	$p_{\cdot 1}$	$p_{\cdot 2}$	$p_{\cdot 3}$	$p_{\cdot 4}$	1

photography can illuminate subtle changes in forest conditions such as decline due to insects or water stress and converse recovery of forests following disturbance. The response design protocols presented also do not address the situation in which the map provides information as a continuous variable. Although many of the basic concepts underlying the good practice recommendations would apply to a continuous variable, the details of the accuracy assessment methodology (cf. Riemann, Wilson, Lister, & Parks, 2010) and area estimation would likely be considerably different from the methods presented herein.

#### 4. Analysis

The analysis protocol specifies the measures to be used to express accuracy and class area as well as the procedures to estimate the selected measures from the sample data. In the context of studies of land change, there are two key objectives of the analysis: 1) accuracy assessment of the change classification, and 2) estimation of area of change. The confusion or error matrix (hereafter noted as the error matrix) plays a central role in meeting both the accuracy assessment and area estimation objectives (Foody, 2013; Stehman, 2013).

##### 4.1. The error matrix

The error matrix is a simple cross-tabulation of the class labels allocated by the classification of the remotely sensed data against the reference data for the sample sites. The error matrix organizes the acquired sample data in a way that summarizes key results and aids the quantification of accuracy and area. The main diagonal of the error matrix highlights correct classifications while the off-diagonal elements show omission and commission errors. The cell entries and marginal values of the error matrix are fundamental to both accuracy assessment and area estimation. Table 4 illustrates a four-class example error matrix of the type often used in studies of land change.

The rows of the error matrix represent the labels shown in a map derived from the classification of the remote sensing data and the columns represent the labels depicted in the reference data. This layout is not a universal requirement and some may wish to reverse the contents of the rows and columns. In the matrix,  $p_{ij}$  represents the proportion of area for the population that has map class  $i$  and reference class  $j$ , where “population” is defined as the full region of interest, and  $p_{ij}$  is therefore the value that would result if a census of the population was obtained (i.e., complete coverage reference classification).

Accuracy parameters derived from a population error matrix of  $q$  classes include overall accuracy

$$O = \sum_{j=1}^q p_{jj} \quad (1)$$

user’s accuracy of class  $i$  (the proportion of the area mapped as class  $i$  that has reference class  $i$ )

$$U_i = p_{ii}/p_i \quad (2)$$

or its complementary measure, commission error of class  $i$ ,  $1 - p_{ii}/p_i$ , and producer’s accuracy of class  $j$  (the proportion of the area of reference class  $j$  that is mapped as class  $j$ ),

$$P_j = p_{jj}/p_j \quad (3)$$

or its complementary measure, omission error of class  $j$ ,  $1 - p_{jj}/p_j$ . A variety of other measures of accuracy has been used in remote sensing (Liu, Frazier, & Kumar, 2007). A commonly used measure is the kappa coefficient of agreement (Congalton & Green, 2009). The problems associated with kappa include but are not limited to: 1) the correction for hypothetical chance agreement produces a measure that is not descriptive of the accuracy a user of the map would encounter (kappa would underestimate the probability that a randomly selected pixel is correctly

classified); 2) the correction for chance agreement used in the common formulation of kappa is based on an assumption of random chance that is not reasonable because it uses the map marginal proportions of area in the definition of chance agreement and these proportions are clearly not simply random; and 3) kappa is highly correlated with overall accuracy so reporting kappa is redundant with overall accuracy.” (Foody, 1992; Liu et al., 2007; Pontius & Millones, 2011; Stehman, 1997). Consistent with the recommendation in Strahler et al. (2006) the use of kappa is strongly discouraged as, despite its widespread use, it actually does not serve a useful role in accuracy assessment or area estimation.

##### 4.2. General principles of estimation for good practice

The analysis protocol is designed to achieve the objectives of estimating accuracy and area from the sample data. Analysis thus requires statistical inference as the underlying scientific support for generalizing from the sample data to the population parameters and for quantifying uncertainty of the sample-based estimators. We recommend design-based inference (Särndal, Swensson, & Wretman, 1992) as the framework within which estimation is conducted. A fundamental tenet of design-based inference is that the specific estimators for accuracy, area, and the variances of these estimators depend on the sampling design implemented; different estimators are appropriate for different sampling designs. Therefore, it is essential that only unbiased or consistent estimators should be used. In practical terms, this means that only formulas for estimating parameters and variances that account for the inclusion probabilities associated with the sampling design implemented should be used. All recommended good practice estimators meet this condition, but the versions of the estimators presented are usually forms where the individual inclusion probabilities do not appear explicitly.

##### 4.3. Estimating accuracy

The cell entries of the population error matrix and the parameters derived from it must be estimated from a sample. Suppose the sample-based estimator of  $p_{ij}$  is denoted as  $\hat{p}_{ij}$ . Once  $\hat{p}_{ij}$  is available for each element of the error matrix, parameters can be estimated by substituting  $\hat{p}_{ij}$  for  $p_{ij}$  in the formulas for the parameters. Accordingly, the error matrix should be reported in terms of these estimated area proportions,  $\hat{p}_{ij}$ , and not in terms of sample counts,  $n_{ij}$ . The specific formula for estimating  $p_{ij}$  depends on the sampling design used. For equal probability sampling designs (e.g., simple random and systematic sampling) and for stratified random sampling in which the strata correspond to the map classes,

$$\hat{p}_{ij} = W_i \frac{n_{ij}}{n_i} \quad (4)$$

where  $W_i$  is the proportion of area mapped as class  $i$ . For simple random and systematic sampling, Eq. (4) is a poststratified estimator of  $p_{ij}$  (Card, 1982) and for these sampling designs the poststratified estimator is recommended because it will have better precision than the estimators commonly used (cf. Stehman & Foody, 2009). Substituting  $\hat{p}_{ij}$  of Eq. (4) into Eqs. (1)–(3) yields estimators of overall, user’s, and producer’s accuracies. These formulas are simpler special cases of a more general estimation approach described in Strahler et al. (2006, Eq. (3.1)).

The sampling variability associated with the accuracy estimates should be quantified by reporting standard errors. The variance estimators are provided below, and taking the square root of the estimated variance results in the standard error of the estimator. For overall accuracy, the estimated variance is

$$\hat{V}(\hat{O}) = \sum_{i=1}^q W_i^2 \hat{U}_i (1 - \hat{U}_i) / (n_i - 1). \quad (5)$$

For user's accuracy of map class  $i$ , the estimated variance is

$$\hat{V}(\hat{U}_i) = \hat{U}_i(1 - \hat{U}_i)/(n_i - 1). \quad (6)$$

For producer's accuracy of reference class  $j = k$ , the estimated variance is

$$\hat{V}(\hat{P}_j) = \frac{1}{\hat{N}_j^2} \left[ \frac{N_j^2 (1 - \hat{P}_j)^2 \hat{U}_j (1 - \hat{U}_j)}{n_j - 1} + \hat{P}_j^2 \sum_{i \neq j} N_i^2 \frac{n_{ij}}{n_i} \left( 1 - \frac{n_{ij}}{n_i} \right) / (n_i - 1) \right] \quad (7)$$

where  $\hat{N}_j = \sum_{i=1}^q \frac{N_i}{n_i} n_{ij}$  is the estimated marginal total number of pixels of reference class  $j$ ,  $N_j$  is the marginal total of map class  $j$  and  $n_j$  is the total number of sample units in map class  $j$ . These are the usual variance estimators applied to the stratified sampling, and the estimators would be viewed as poststratified variance estimators for simple random and systematic sampling. For systematic sampling, the variance estimators are approximations that usually result in overestimation of variance. These variance estimators are also based on assumptions that the assessment unit for the response design is a pixel and each pixel has a hard classification for the map and a hard classification for the reference data. The variance estimators would not apply to a polygon assessment unit or to a mixed pixel situation.

#### 4.4. Estimating area

The error matrix also provides the basis for estimating the area of classes such as those representing change and no-change. The population error matrix (Table 4) provides two different approaches for estimating the proportion of area. Suppose we are interested in estimating the proportion of area of class  $k$ . The row and column totals are the sums of the  $p_{ij}$  values in the respective rows and columns. Thus, the row total  $p_{\cdot k}$  represents the proportion of area mapped as class  $k$  (e.g., if  $k$  is a change class such as forest loss then  $p_{\cdot k}$  is the proportion of area mapped as forest loss) and the column total  $p_{\cdot k}$  represents the proportion of area of class  $k$  as determined from the reference classification (e.g.,  $p_{\cdot k}$  would be the proportion of area of forest loss as determined from the reference classification).

The two area proportion parameters for class  $k$  (i.e.,  $p_{\cdot k}$  and  $p_{\cdot k}$ ) are unlikely to have the same value, so a decision arises as to which parameter should be the focus. Once a change map is complete,  $p_{\cdot k}$  is known, but because the reference classification is available only for a sample,  $p_{\cdot k}$  must be estimated from the sample. Consequently, the need to estimate  $p_{\cdot k}$  introduces uncertainty in the form of sampling variability, whereas  $p_{\cdot k}$  is not subject to sampling variability (Stehman, 2005). The map-based parameter  $p_{\cdot k}$  is known with certainty but likely biased because of classification error. Conversely,  $p_{\cdot k}$  is determined from the reference classification. Therefore,  $p_{\cdot k}$  should have smaller bias than  $p_{\cdot k}$  (i.e., the bias attributable to reference data error is smaller than the bias attributable to map classification error). The "good practice" guidelines are founded on the premise that the reference classification is superior in quality relative to the map classification and that the sampling design implemented yields estimates with small standard errors. Consequently, we recommend that area estimation should be based on  $p_{\cdot k}$ , the proportion of area derived from the reference classification.

A variety of estimators has been proposed for estimating  $p_{\cdot k}$  from the error matrix. For any sampling design and response design leading to an estimated error matrix with  $p_{ij}$  in terms of proportion of area, a direct estimator of the proportion of area of class  $k$  is

$$\hat{p}_{\cdot k} = \sum_{i=1}^q \hat{p}_{ik}. \quad (8)$$

This estimator is simply the sum of the estimated area proportions of class  $k$  as determined from the reference classification (i.e., the sum of column  $k$  of the estimated error matrix). If the sampling design is simple

random, systematic, or stratified random with the map classes defined as the strata, Eq. (8) would be computed using  $\hat{p}_{ij} = W_i \frac{n_{ij}}{n_i}$  leading to the often used special case estimator

$$\hat{p}_{\cdot k} = \sum_{i=1}^q W_i \frac{n_{ik}}{n_i}. \quad (9)$$

This estimator is a poststratified estimator for simple random and systematic sampling, and it is the direct stratified estimator of  $p_{\cdot k}$  for stratified random sampling when the map classes are the strata. For these sampling designs, the stratified estimator (Eq. (9)) generally has better precision than a variety of alternative estimators of area (Stehman, 2013) and consequently the stratified estimator is recommended.

For the stratified estimator of proportion of area (Eq. (9)), the standard error is estimated by

$$S(\hat{p}_{\cdot k}) = \sqrt{\sum_i W_i^2 \frac{n_{ik} \left( 1 - \frac{n_{ik}}{n_i} \right)}{n_i - 1}} = \sqrt{\sum_i W_i \frac{\hat{p}_{ik} - \hat{p}_{\cdot k}^2}{n_i - 1}} \quad (10)$$

where  $n_{ik}$  is the sample count at cell  $(i, k)$  in the error matrix,  $W_i$  is the area proportion of map class  $i$ ,  $\hat{p}_{ik} = W_i \frac{n_{ik}}{n_i}$  and the summation is over the  $q$  classes. For systematic sampling, Eq. (10) is an approximation that is typically an overestimate for the actual standard error of systematic sampling. The estimated area of class  $k$  is  $\hat{A}_k = A \times \hat{p}_{\cdot k}$ , where  $A$  is the total map area. The standard error of the estimated area is given by

$$S(\hat{A}_k) = A \times S(\hat{p}_{\cdot k}). \quad (11)$$

An approximate 95% confidence interval is obtained as  $\hat{A}_k \pm 1.96 \times S(\hat{A}_k)$ .

### 5. Example of good practices: estimating area and assessing accuracy of forest change

The following hypothetical example illustrates the workflow of assessing accuracy of a forest change map and estimating area. Consider a change map for 2000 to 2010 consisting of two change classes and two stable classes: deforestation, forest gain, stable forest and stable non-forest. The map was produced by supervised classification of data from Landsat ETM+ with the objective of estimating the gross rates of forest loss and gain. The first step in the assessment was to visually inspect the change map and identify obvious errors by comparing the classified results to the Landsat data of 2000 and 2010. Misclassified regions were relabeled before proceeding to the rigorous evaluation of the map. After obvious errors were removed, the areas of the map classes were 200,000 Landsat pixels (18,000 ha) of deforestation, 150,000 pixels (13,500 ha) of forest gain, 3,200,000 pixels (288,000 ha) of stable forest, and 6,450,000 pixels (580,500 ha) of stable non-forest. The two change classes thus occupy 3.5% of the total map area. The accuracy assessment was designed for the objectives of estimating overall and class-specific accuracies, areas of the individual classes (as determined by the reference classification), and confidence intervals for each accuracy and area parameter. The spatial assessment unit in this example is a Landsat pixel (30 m  $\times$  30 m).

#### 5.1. Sampling design

A stratified random sampling design with the four map classes as strata adheres to the recommended practices outlined in Section 2 and satisfies the accuracy assessment and area estimation objectives. In the next two subsections, we present sample size and sample allocation planning calculations for the stratified design. Sample size planning is an inexact science because it is dependent on accuracy and area information that must be speculative prior to conducting the actual accuracy

assessment. Nevertheless, these planning calculations can provide informative insight into the choices of sample size and sample allocation to strata.

5.1.1. Determining the sample size

For simple random sampling and targeting overall accuracy as the estimation objective, Cochran (1977, Eq. (4.2)) suggests using a sample size of

$$n = \frac{z^2 O(1-O)}{d^2} \tag{12}$$

where  $O$  is the overall accuracy expressed as a proportion,  $z$  is a percentile from the standard normal distribution ( $z = 1.96$  for a 95% confidence interval,  $z = 1.645$  for a 90% confidence interval), and  $d$  is the desired half-width of the confidence interval of  $O$ . Eq. (12) provides a starting point for assessing sample size for the limited scope of estimating overall accuracy.

For stratified random sampling, Cochran (1977, Eq. (5.25)) provides the following sample size formula (the cost of sampling each stratum is assumed the same):

$$n = \frac{(\sum W_i S_i)^2}{[s(\hat{O})]^2 + (1/N)\sum W_i S_i^2} \approx \left(\frac{\sum W_i S_i}{s(\hat{O})}\right)^2 \tag{13}$$

where  $N$  = number of units in the ROI,  $s(\hat{O})$  is the standard error of the estimated overall accuracy that we would like to achieve,  $W_i$  is the mapped proportion of area of class  $i$ , and  $S_i$  is the standard deviation of stratum  $i$ ,  $S_i = \sqrt{U_i(1-U_i)}$  (Cochran, 1977, Eq. (5.55)). Because  $N$  is typically large (e.g., over 10 million pixels in this example), the second term in the denominator of Eq. (13) can be ignored. We specify a target standard error for overall accuracy of 0.01. Suppose from past experience with similar change mapping efforts we know that errors of commission are relatively common for the change classes while the stable classes are more accurate (e.g., Olofsson et al., 2010, 2011). Consequently, we conjecture that user's accuracies of the two change classes will be 0.70 for deforestation and 0.60 for forest gain, and user's accuracies of the stable classes will be 0.90 for stable forest and 0.95 for stable non-forest. The resulting sample size from Eq. (13) is  $n = 641$ . These sample size calculations should be repeated for a variety of choices of  $s(\hat{O})$  and  $U_i$  before reaching a final decision.

5.1.2. Determine sample allocation to strata

Once the overall sample size is chosen, we determine the allocation of the sample to strata. It is important that the sample size allocation results in precise estimates of accuracy and area. Stehman (2012) identifies four different approaches to sample allocation: proportional, equal, optimal and power allocation. In proportional allocation, the sample size per map class is proportional to the relative area of the map class. In this example, and which is usually the case when mapping land change, the mapped areas of change are small relative to other classes so proportional allocation will lead to small sample sizes in the rare classes (unless  $n$  is very large) and imprecise estimates of user's accuracy for these rare classes. Allocating an equal sample size to all strata targets estimation of user's accuracy of each map class but equal allocation is not optimized for estimating area and overall accuracy. Neyman optimal allocation (Cochran, 1977) can be used to minimize the variance of the estimator of overall accuracy or the estimator of area, but optimal allocation becomes difficult to implement when multiple estimation objectives are of interest as will be the case when estimating accuracy and area of several land-cover classes or land-cover change types.

We suggest the following simplified approach to sample size allocation. Allocate a sample size of 50–100 for each change strata using the variance estimator for user's accuracy (Eq. (6)) to decide the sample

size needed to achieve certain standard errors for the assumed estimated user's accuracy for that class. A small overall sample size might allow for only 50 sample units per rare class stratum. Suppose that  $n-r$  sample units remain after a sample size of  $r$  units has been allocated to the rare class strata. The sample size of  $n-r$  is then allocated proportionally to the area of each remaining stratum. The anticipated estimated variances can then be computed (based on the sample size allocation) for user's and overall accuracy and area using Eqs. (5), (6) and (10). The sample size allocation process can be iterated until an allocation is found that yields satisfactory anticipated standard errors for the key accuracy and area estimates. The effect of the choice of sample allocation will be observed in the standard errors of the estimates, however, a poor allocation of sample size to strata will not result in biased estimators.

In this example, we know the mapped areas of the four map classes ( $W_i$ ), we have conjectured values of user's accuracies and standard errors of the strata, and we have estimated a total sample size of 641 (Table 5). The resulting sample sizes for proportional and equal allocation are shown in Table 5. As described above, neither of these is optimal and we want to find a compromise between the two. We start by allocating 100 sample units each to the change classes and then allocate the remainder of the sample size proportionally to the stable classes. This gives the allocation in column "Alloc1". Since the recommendation is to allocate between 50 and 100 sample units in the change strata, we introduce two additional allocations with 75 and 50 sample units in the change strata, respectively ("Alloc2" and "Alloc3"). To determine which of these allocations to use, we need to examine the standard errors of the estimated user's accuracy, estimated overall accuracy, and estimated areas using Eqs. (5), (6) and (10).

It is necessary to speculate the outcome of the accuracy assessment to compute the anticipated standard errors for each sample allocation considered. The hypothesized error matrix in Table 6 reflects the anticipated outcome that the change classes will be rare and have lower class-specific accuracies than the two stable classes. The population error matrix was also constructed to yield the hypothesized accuracies input into the sample size planning calculations of the previous section. When creating the hypothesized error matrix used for sample size and sample allocation planning, we should draw upon any past experience for insight into the accuracy of the map to be produced.

Table 7 shows the standard errors of the user's and overall accuracies and estimated areas of both deforestation and stable forest for each of the five sample allocations in Table 5 and the hypothetical population error matrix of Table 6. No single allocation is best for all estimation objectives, so a choice among competing objectives is necessary. The emphasis on prioritizing objectives during the planning stage (Section 2) becomes particularly relevant to the decision of sample allocation because different allocations favor different estimation objectives. For example, equal allocation gives the smallest standard error of the user's accuracy of deforestation but a high standard error of the estimated area of deforestation. Proportional allocation will result in smaller standard errors of overall accuracy and area of stable forest but the standard error for estimated user's accuracy of deforestation is two to four times larger than the corresponding standard errors for other sample allocations. In this case, "Alloc1–3" provide allocations that generate relatively small standard errors for the different estimates. We will choose "Alloc2" with 75 sample units in the two change classes.

Table 5

Information needed to decide allocation of sample size to strata. The information includes the mapped area proportions ( $W_i$ ), conjectured values of user's accuracies ( $U_i$ ) and standard deviations ( $S_i$ ) of the strata. Columns 5–9 contain five different allocations.

Strata (i)	$W_i$	$U_i$	$S_i$	Equal	Alloc1	Alloc2	Alloc3	Prop
1 Deforestation	0.020	0.700	0.458	160	100	75	50	13
2 Forest gain	0.015	0.600	0.490	160	100	75	50	10
3 Stable forest	0.320	0.900	0.300	160	149	165	182	205
4 Stable non-forest	0.645	0.950	0.218	160	292	325	358	413

**Table 6**  
Hypothetical population error matrix expressed in terms of proportion of area (see Section 4) used for sample size and sample allocation planning calculations.

		Reference					
		Deforestation	Forest gain	Stable forest	Stable non-forest	Total ( $W_i$ )	$U_i$
Map	Deforestation	0.014	0	0.003	0.003	0.020	0.70
	Forest gain	0	0.009	0.003	0.003	0.015	0.60
	Stable forest	0.002	0	0.288	0.030	0.320	0.90
	Stable non-forest	0.004	0.002	0.025	0.614	0.645	0.95
	Total	0.020	0.011	0.319	0.650	1	

## 5.2. Estimating accuracy, area and confidence intervals

To create the reference classification for labeling each sample unit, a combination of Landsat data from the USGS open archive together with GoogleEarth™ provides a source of cost free reference data. Our hypothetical map was produced using Landsat, and the good practice recommendations stipulate that if using the same data for creation of both the map and reference classifications, the process of creating the latter should be of higher quality than the map-making process. The process of labeling the sample units thus has to be more accurate than supervised classification. A manual inspection by three analysts of each of the sample units using a set of Landsat images together with GoogleEarth™ imagery acquired around the same time as the images used to make the map is assumed to be a more accurate process than supervised classification. The error matrix resulting from this response design and sample is presented in terms of the sample counts displayed in Table 8, and the computations for the accuracy and area estimates are detailed in the following two subsections.

### 5.2.1. Estimating accuracy

Because the sampling design is stratified random using the map classes as strata, the cell entries of the error matrix are estimated using Eq. (4).

We can now estimate user's accuracy  $\hat{U}_i = \frac{\hat{p}_{ii}}{\hat{p}_{i.}}$ ; producer's accuracy  $\hat{P}_j = \frac{\hat{p}_{jj}}{\hat{p}_{.j}}$ ; and overall accuracy  $\hat{O} = \sum_{j=1}^q \hat{p}_{jj}$  using the estimated area proportions. Variances for these accuracy measures are estimated using Eqs. (5)–(7). 95% confidence intervals are estimated as  $\pm 1.96 \sqrt{\hat{V}(\hat{U}_i)}$  (replace  $\hat{U}_i$  with  $\hat{P}_j$  and  $\hat{O}$  for the producer's and overall accuracies). In this case, the estimated user's accuracy ( $\pm 95\%$  confidence interval) is  $0.88 \pm 0.07$  for deforestation,  $0.73 \pm 0.10$  for forest gain,  $0.93 \pm 0.04$  for stable forest, and  $0.96 \pm 0.02$  for stable non-forest. The estimated producer's accuracy is  $0.75 \pm 0.21$  for deforestation,  $0.85 \pm 0.23$  for forest gain,  $0.93 \pm 0.03$  for stable forest, and  $0.96 \pm 0.01$  for stable non-forest. The estimated overall accuracy is  $0.95 \pm 0.02$ .

### 5.2.2. Estimating area and uncertainty

The next step is to use the estimated area proportions in Table 9 to estimate the area of each class. The row totals of the error matrix in

**Table 7**  
Standard errors of selected accuracy and area estimates for different sample size allocations to strata (Table 5) and the hypothetical population error matrix (Table 6). Standard errors are shown for estimated overall accuracy, estimated user's accuracy for the rare class deforestation ( $i = 1$ ) and the common class stable forest ( $i = 3$ ), and estimated area (in units of hectares) of deforestation and area of stable forest.

Allocation	$s(\hat{O})$	$s(\hat{U}_1)$	$s(\hat{U}_3)$	$s(\hat{A}_1)$	$s(\hat{A}_3)$
Equal	0.013	0.036	0.024	4035	11,306
Alloc1	0.011	0.046	0.025	3307	9744
Alloc2	0.011	0.053	0.023	3138	9270
Alloc3	0.010	0.065	0.022	3125	8860
Proportional	0.010	0.132	0.021	3600	8614

Table 9 give the mapped area proportions (which are also given by  $W_i$ ) while the column totals give the estimated area proportions according to the reference data. Multiplying the latter by the total map area gives the stratified area estimate of each class according to the reference data. For example, the estimated area of deforestation according to the reference data is  $\hat{A}_1 = \hat{p}_{.1} \times A_{tot} = 0.024 \times 10,000,000 \text{ pixels} = 235,086 \text{ pixels} = 21,158 \text{ ha}$ . The mapped area of deforestation ( $A_{m,1}$ ) of 200,000 pixels was thus underestimated by 35,086 pixels or 3158 ha.

The second step is to estimate a confidence interval for the area of each class. From Eq. (10),  $S(\hat{p}_{.1}) = 0.0035$  and the standard error for the estimated area of forest loss is  $S(\hat{A}_1) = S(\hat{p}_{.1}) \times A_{tot} = 0.0035 \times 10,000,000 = 34,097 \text{ pixels}$ . The margin of error of the confidence interval is  $1.96 \times 34,097 = 68,418 \text{ pixels} = 6158 \text{ ha}$ . We have thus estimated the area of deforestation with a 95% confidence interval:  $21,158 \pm 6158 \text{ ha}$ . The area estimate with a 95% confidence interval of the forest gain class is  $11,686 \pm 3756 \text{ ha}$ ; stable forest is  $285,770 \pm 15,510 \text{ ha}$  and stable non-forest  $581,386 \pm 16,282 \text{ ha}$ .

This example has illustrated the workflow of assessing accuracy, and estimating area and confidence intervals of area of the classes of a change map. While this is fairly straightforward once the error matrix has been constructed, the example highlights the need to consider different objectives when designing the sample.

A tool for estimating unbiased accuracy measures and areas with 95% confidence intervals can be downloaded from [www.people.bu.edu/olofsson/](http://www.people.bu.edu/olofsson/) (click 'Research' > 'Accuracy/Uncertainty'). The tool is implemented in Matlab™.

## 6. Summary

Conducting an accuracy assessment of a land change map serves multiple purposes. In addition to the obvious purpose of quantifying the accuracy of the map, the reference sample serves as the basis of estimates of area of each class where area is defined by the reference classification. The accuracy assessment sample data also contribute to estimates of uncertainty of the area estimates. Without an accuracy assessment, there is no way to communicate map quality in a quantitative and meaningful fashion. We acknowledge that there is no singular "best" approach and the recommendations provided do not preclude the existence of other acceptable practices. However, by following the "good practice" recommendations presented by this paper, scientific credibility of the accuracy and area estimates is ensured. The "good practice" recommendations are summarized as follows, organized by the three major components of the accuracy assessment methodology, the sampling design, response design, and analysis.

### 6.1. General

- Visually inspect the map and correct obvious errors before conducting the accuracy assessment.
- Accuracy and area estimates will be determined from a classification (i.e., the reference classification) that is of higher quality than the land change map being evaluated.
- A sampling approach is needed because the cost of obtaining the reference classification for the entire region of interest will be prohibitive.

**Table 8**

Description of sample data as an error matrix of sample counts,  $n_{ij}$  (see Table 9 for recommended estimated error matrix used to report accuracy results).

		Reference				Total	$A_{m,i}$ [pixels]	$W_i$
		Deforestation	Forest gain	Stable forest	Stable non-forest			
Map	Deforestation	66	0	5	4	75	200,000	0.020
	Forest gain	0	55	8	12	75	150,000	0.015
	Stable forest	1	0	153	11	165	3,200,000	0.320
	Stable non-forest	2	1	9	313	325	6,450,000	0.645
	Total	69	56	175	340	640	10,000,000	1

- The sample used for accuracy assessment and area estimation is separate from (independent of) the data used to train or develop the classification.

6.2. Sampling design

- Implement a probability sampling design to provide a rigorous foundation via design-based sampling inference.
- Document and quantify any deviations from the probability sampling protocol.
- Choose a sampling design on the basis of specified accuracy objectives and prioritized desirable design criteria.
- Sampling design guidelines.
  - Stratify by map class to reduce standard errors of class-specific accuracy estimates.
  - If resources are adequate, stratify by subregions to reduce standard errors of subregion-specific estimates.
  - Use cluster sampling if it provides a substantial cost savings or if the objectives require a cluster unit for the assessment.
  - Both simple random and systemic selection protocols are acceptable options.
- The recommended allocation of sample size to strata (assuming the map classes are the strata) is to increase the sample size for rare change classes to achieve an acceptable standard error for estimated user's accuracies and to allocate the remaining sample size roughly proportional to the area occupied by the common classes.
- Use sample size and optimal allocation planning calculations as a guide to decisions on total sample size and sample allocation.
- Evaluate the potential outcome of sample size and sample allocation decisions on the standard errors of accuracy and area estimates for hypothetical error matrices based on the anticipated accuracy of the map.
- Stratified random sampling using the map classification to define strata is a simple, but generally applicable design that will typically satisfy most accuracy and area estimation objectives and desirable design criteria.

6.3. Response design

- Reference data should be of higher quality than the data used for creating the map, or if using the same source, the process of creating the reference classification should be more accurate than the process of creating the map.

- High overhead cost may eliminate field visits as a source of reference data.
- The reference data should provide sufficient temporal representation consistent with the change period of the map.
- Data from the Landsat open archive in combination with high spatial resolution imagery provide a low-cost and often useful source of reference data (national photograph archives, satellite photo archives (e.g., Kompsat), and the collections available through Google Earth™ are possible high resolution imagery sources).
- Specify protocols for accounting for uncertainty in assigning the reference classifications.
- Assign each sample unit a primary and secondary label (secondary not required if there is highly confidence in the primary label).
- Include an interpreter specified confidence for each reference label (e.g., high, medium, or low confidence).
- Implement protocols to ensure consistency among individual interpreters or teams of interpreters.
- Specify a protocol for defining agreement between the map and reference classifications that will lead to an error matrix expressed in terms of proportion of area.

6.4. Analysis

- Report the error matrix in terms of estimated area proportions.
- Report the area (or proportion of area) of each class as determined from the map.
- Report user's accuracy (or commission error), producer's accuracy (or omission error), and overall accuracy (Eqs. (1)–(3)).
- Avoid use of the kappa coefficient of agreement for reporting accuracy of land change maps.
- Estimate the area of each class according to the classification determined from the reference data.
- Use estimators of accuracy and area that are unbiased or consistent.
- For simple random, systematic, and stratified random sampling when the map classes are defined as strata, use stratified estimators of accuracy (Eqs. (5)–(7)) and a stratified estimator of area (Eq. (9)).
- Quantify sampling variability of the accuracy and area estimates by reporting standard errors or confidence intervals.
- Use design-based inference to define estimator properties and to quantify uncertainty.
- Assess the impact of reference data uncertainty on the accuracy and area estimates.

**Table 9**

The error matrix in Table 8 populated by estimated proportions of area.

		Reference				Total ( $W_i$ )	$A_{m,i}$ [pixels]
		Deforestation	Forest gain	Stable forest	Stable non-forest		
Map	Deforestation	0.0176	0	0.0013	0.0011	0.020	200,000
	Forest gain	0	0.0110	0.0016	0.0024	0.015	150,000
	Stable forest	0.0019	0	0.2967	0.0213	0.320	3,200,000
	Stable non-forest	0.0040	0.0020	0.0179	0.6212	0.645	6,450,000
	Total	0.0235	0.0130	0.3175	0.6460	1	10,000,000



The recommendations provided are intended to serve as guidelines for choosing from among options of sampling design, response design, and analysis that will yield rigorous and defensible accuracy and area estimates. But good practice is not static. As improvements in technology become available and new methods are developed, good practice recommendations will evolve over time. Also, as practical experience accumulates with using new technology and methodologies, good practice recommendations will be further amended to provide even more efficient yet still rigorous methods to estimate accuracy and area of land change.

## Acknowledgments

This research was funded by the USGS Award Support for SilvaCarbon and NASA through its support for the Carbon Monitoring System to Boston University, and NASA Grant Number NNX13AP48G to State University of New York. We acknowledge the European Space Agency (ESA) and NASA for their support to GOF-C-GOLD and the CEOS working group of calibration and validation. We thank the anonymous reviewers for the comments that helped improve the manuscript.

## References

- Achard, F., Eva, H., Stibig, H. -J., Mayaux, P., Gallego, J., Richards, T., et al. (2002). Determination of deforestation rates of the world's humid tropical forests. *Science*, 297, 999–1002.
- Ahlqvist, O. (2008). In search of classification that supports the dynamics of science: The FAO Land Cover Classification System and proposed modifications. *Environment and Planning B: Planning and Design*, 35, 169–186.
- Baker, B.A., Warner, T. A., Conley, J. F., & McNeil, B. E. (2013). Does spatial resolution matter? A multi-scale comparison of object-based and pixel-based methods for detecting change associated with gas well drilling operations. *International Journal of Remote Sensing*, 34, 1633–1651.
- Binaghi, E., Brivio, P. A., Ghezzi, P., & Rampini, A. (1999). A fuzzy set-based accuracy assessment of soft classification. *Pattern Recognition Letters*, 20, 935–948.
- Cakir, H. I., Khorram, S., & Nelson, S. A. C. (2006). Correspondence analysis for detecting land cover change. *Remote Sensing of Environment*, 102, 306–317.
- Card, D. H. (1982). Using map category marginal frequencies to improve estimates of thematic map accuracy. *Photogrammetric Engineering and Remote Sensing*, 49, 431–439.
- Cochran, W. G. (1977). *Sampling techniques* (3rd ed.). New York: John Wiley & Sons.
- Cohen, W. B., Yang, Z., & Kennedy, R. (2010). Detecting trends in forest disturbance and recovery using yearly Landsat time series: 2. TimeSync – Tools for calibration and validation. *Remote Sensing of Environment*, 114, 2911–2924.
- Comber, A. J., Wadsworth, R. A., & Fisher, P. F. (2008). Using semantics to clarify the conceptual confusion between land cover and land use: The example of 'forest'. *Journal of Land Use Science*, 3, 185–198.
- Congalton, R., & Green, K. (2009). *Assessing the accuracy of remotely sensed data: Principles and practices* (2nd ed.). Boca Raton: CRC/Taylor & Francis.
- de Sy, V., Herold, M., Achard, F., Asner, G. P., Held, A., Kellndorfer, J., et al. (2012). Synergies of multiple remote sensing data sources for REDD+ monitoring. *Current Opinion in Environmental Sustainability*, 4, 696–706.
- DeFries, R., Achard, F., Brown, S., Herold, M., Murdiyasar, D., Schlamadinger, B., et al. (2007). Earth observations for estimating greenhouse gas emissions from deforestation in developing countries. *Environmental Science and Policy*, 10, 385–394.
- DeFries, R., Houghton, R. A., Hansen, M., Field, C., Skole, D. L., & Townshend, J. (2002). Carbon emissions from tropical deforestation and regrowth based on satellite observations for the 1980s and 90s. *Proceedings of the National Academy of Sciences*, 99, 14256–14261.
- Drummond, M.A., & Loveland, T. R. (2010). Land-use pressure and a transition to forest-cover loss in the eastern United States. *BioScience*, 60, 286–298.
- Duro, D. C., Franklin, S. E., & Duba, M. G. (2012). A comparison of pixel-based and object-based image analysis with selected machine learning algorithms for the classification of agricultural landscapes using SPOT-5 HRG imagery. *Remote Sensing of Environment*, 118, 259–272.
- Falkowski, M. J., Wulder, M.A., White, J. C., & Gillis, M.D. (2009). Supporting large-area, sample-based forest inventories with very high spatial resolution satellite imagery. *Progress in Physical Geography*, 33, 403–423.
- FAO (2010). *Global forest resources assessment 2010*. Food and Agriculture Organization of the United Nations.
- Foody, G. M. (1992). On the compensation for chance agreement in image classification accuracy assessment. *Photogrammetric Engineering and Remote Sensing*, 58, 1459–1460.
- Foody, G. M. (1996). Approaches for the production and evaluation of fuzzy land cover classifications from remotely sensed data. *International Journal of Remote Sensing*, 17, 1317–1340.
- Foody, G. M. (2002). Status of land cover classification accuracy assessment. *Remote Sensing of Environment*, 80, 185–201.
- Foody, G. M. (2010). Assessing the accuracy of land cover change with imperfect ground reference data. *Remote Sensing of Environment*, 114, 2271–2285.
- Foody, G. M. (2013). Ground reference data error and the mis-estimation of the area of land cover change as a function of its abundance. *Remote Sensing Letters*, 4, 783–792.
- Foody, G. M., & Boyd, D. S. (2013). Using volunteered data in land cover map validation: Mapping West African forests. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 6, 1305–1312.
- Foody, G. M., Campbell, N. A., Trodd, N. M., & Wood, T. F. (1992). Derivation and applications of probabilistic measures of class membership from the maximum likelihood classification. *Photogrammetric Engineering and Remote Sensing*, 58, 1335–1341.
- Gallego, F. J. (2012). The efficiency of sampling very high resolution images for area estimation in the European Union. *International Journal of Remote Sensing*, 33, 1868–1880.
- GOF-C-GOLD (2011). A sourcebook of methods and procedures for monitoring and reporting anthropogenic greenhouse gas emissions and removals caused by deforestation, gains and losses of carbon stocks in forests remaining forests, and forestation. *GOF-C-GOLD Report version COP17-1*, (GOF-C-GOLD Project Office, Natural Resources Canada, Alberta, Canada).
- Gómez, C., White, J. C., & Wulder, M.A. (2011). Characterizing the state and processes of change in a dynamic forest environment using hierarchical spatio-temporal segmentation. *Remote Sensing of Environment*, 115, 1665–1679.
- Gopal, S., & Woodcock, C. (1994). Theory and methods for accuracy assessment of thematic maps using fuzzy sets. *Photogrammetric Engineering and Remote Sensing*, 60, 181–188.
- Grassi, G., Monni, S., Federici, S., Achard, F., & Mollicone, D. (2008). Applying the conservativeness principle to REDD to deal with the uncertainties of the estimates. *Environmental Research Letters*, 3, 3.
- Hansen, M. C., Stehman, S. V., & Potapov, P. V. (2010). Quantification of global gross forest cover loss. *Proceedings of the National Academy of Sciences*, 107, 8650–8655.
- He, Y. H., Franklin, S. E., Guo, X. L., & Stenhouse, G. B. (2011). Object-orientated classification of multi-resolution images for the extraction of narrow linear forest disturbance. *Remote Sensing Letters*, 2, 147–155.
- Huang, C., Goward, S. N., Masek, J. G., Thomas, N., Zhu, Z., & Vogelmann, J. E. (2010). An automated approach for reconstructing recent forest disturbance history using dense Landsat time series stacks. *Remote Sensing of Environment*, 114, 183–198.
- Hyypää, J., Hyypää, H., Inkinen, M., Engdahl, M., Linko, S., & Zhu, Y. H. (2000). Accuracy comparison of various remote sensing data sources in the retrieval of forest stand attributes. *Forest Ecology and Management*, 128, 109–120.
- Iwao, K., Nishida, K., Kinoshita, T., & Yamagata, Y. (2006). Validating land cover maps with Degree Confluence Project information. *Geophysical Research Letters*, 33 (L23404).
- Jeon, S. B., Olofsson, P., & Woodcock, C. E. (2013). Land use change in New England: A reversal of the forest transition. *Journal of Land Use Science*. <http://dx.doi.org/10.1080/1747423X.2012.754962>.
- Johnson, B.A. (2013). High-resolution urban land-cover classification using a competitive multi-scale object-based approach. *Remote Sensing Letters*, 4, 131–140.
- Kelly, M., Estes, J. E., & Knight, K. A. (1999). Image interpretation keys for validation of global land-cover data sets. *Photogrammetric Engineering & Remote Sensing*, 65, 1041–1050.
- Kennedy, R., Yang, Z., & Cohen, W. B. (2010). Detecting trends in forest disturbance and recovery using yearly Landsat time series: 1. LandTrendr – Temporal segmentation algorithms. *Remote Sensing of Environment*, 114, 2897–2910.
- Knight, J. F., & Lunetta, R. S. (2003). An experimental assessment of minimum mapping unit size. *IEEE Transactions on Geoscience and Remote Sensing*, 40, 2132–2134.
- Kurz, W. A. (2010). An ecosystem context for global gross forest cover loss estimates. *Proceedings of the National Academy of Sciences*, 107, 9025–9026.
- Lewis, H. G., & Brown, M. (2001). A generalized confusion matrix for assessing area estimates from remotely sensed data. *International Journal of Remote Sensing*, 22, 3223–3235.
- Lindberg, E., Olofsson, K., Holmgren, J., & Olsson, H. (2012). Estimation of 3D vegetation structure from waveform and discrete return airborne laser scanning data. *Remote Sensing of Environment*, 118, 151–161.
- Liu, C., Frazier, P., & Kumar, L. (2007). Comparative assessment of the measures of thematic classification accuracy. *Remote Sensing of Environment*, 107, 606–616.
- Mayaux, P., Eva, H., Gallego, J., Strahler, A. H., Herold, M., Agrawal, S., et al. (2006). Validation of the Global Land Cover 2000 map. *IEEE Transactions on Geoscience and Remote Sensing*, 44, 1728–1739.
- McRoberts, R. E. (2011). Satellite image-based maps: Scientific inference or pretty pictures? *Remote Sensing of Environment*, 115, 715–724.
- Olofsson, P., Foody, G. M., Stehman, S. V., & Woodcock, C. E. (2013). Making better use of accuracy data in land change studies: Estimating accuracy and area and quantifying uncertainty using stratified estimation. *Remote Sensing of Environment*, 129, 122–131.
- Olofsson, P., Kuemmerle, T., Griffiths, P., Knorn, J., Baccini, A., Gancz, V., et al. (2011). Carbon implications of forest restitution in post-socialist Romania. *Environmental Research Letters*, 6, 045202.
- Olofsson, P., Stehman, S. V., Woodcock, C. E., Sulla-Menashe, D., Sibley, A.M., Newell, J.D., et al. (2012). A global land cover validation dataset, I: Fundamental design principles. *International Journal of Remote Sensing*, 33, 5768–5788.
- Olofsson, P., Torchinava, P., Woodcock, C. E., Baccini, A., Houghton, R. A., Ozdogan, M., et al. (2010). Implications of land use change on the national terrestrial carbon budget of Georgia. *Carbon Balance and Management*, 5, 4.
- Pontius, R. G. (2000). Quantification error versus location error in comparison of categorical maps. *Photogrammetric Engineering & Remote Sensing*, 66, 1011–1016.
- Pontius, R. G., & Lippitt, C. D. (2006). Can error explain map differences over time? *Cartography and Geographic Information Science*, 33, 159–171.
- Pontius, R. G., & Millones, M. (2011). Death to kappa: Birth of quantity disagreement and allocation disagreement for accuracy assessment. *International Journal of Remote Sensing*, 32, 4407–4429.
- Powell, R., Matzke, N., de Souza, C., Clark, M., Numata, I., Hess, L., et al. (2004). Sources of error in accuracy assessment of thematic land-cover maps in the Brazilian Amazon. *Remote Sensing of Environment*, 90, 221–234.

- Pratihast, A. K., Herold, M., de Sy, V., Murdiyarto, D., & Skutsch, M. (2013). Linking community-based and national REDD+ monitoring: A review of the potential. *Carbon Management*, 4, 91–104.
- Riemann, R., Wilson, B. T., Lister, A., & Parks, S. (2010). An effective assessment protocol for continuous geospatial datasets of forest characteristics using USFS Forest Inventory and Analysis (FIA) data. *Remote Sensing of Environment*, 114, 2337–2352.
- Romijn, J. E., Herold, M., Kooistra, L., Murdiyarto, D., & Verchot, L. (2012). Assessing capacities of non-Annex I countries for national forest monitoring in the context of REDD+. *Environmental Science and Policy*, 20, 33–48.
- Sanz-Sanchez, M., Herold, M., & Penman, J. (2013). REDD+ related forest monitoring remains key issue: A report following the recent UN climate convention in Doha. *Carbon Management*, 4, 125–127.
- Särndal, C., Swensson, B., & Wretman, J. (1992). *Model assisted survey sampling*. New York: Springer.
- Saura, S. (2002). Effects of minimum mapping unit on land cover data spatial configuration and composition. *International Journal of Remote Sensing*, 23, 4853–4880.
- Scepan, J. (1999). Thematic validation of high-resolution global land-cover data sets. *Photogrammetric Engineering & Remote Sensing*, 65, 1051–1060.
- Schroeder, T. A., Wulder, M.A., Healey, S. P., & Moisen, G. G. (2011). Mapping wildfire and clearcut harvest disturbances in boreal forests with Landsat time series data. *Remote Sensing of Environment*, 115, 1421–1433.
- Skirvin, S. M., Kepner, W. G., Marsh, S. E., Drake, S. E., Maingi, J. K., Edmonds, C. M., et al. (2004). Assessing the accuracy of satellite-derived land-cover classification using historical aerial photography, digital orthophoto quadrangles, and airborne video data. In R. S. Lunetta, & J. G. Lyon (Eds.), *Remote sensing and GIS accuracy assessment*. Boca Raton: CRC Press.
- 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.
- Stehman, S. V. (2001). Statistical rigor and practical utility in thematic map accuracy. *Photogrammetric Engineering and Remote Sensing*, 67, 727–734.
- Stehman, S. V. (2005). Comparing estimators of gross change derived from complete coverage mapping versus statistical sampling of remotely sensed data. *Remote Sensing of Environment*, 96, 466–474.
- Stehman, S. V. (2009). Sampling designs for accuracy assessment of land cover. *International Journal of Remote Sensing*, 30, 5243–5272.
- Stehman, S. V. (2012). Impact of sample size allocation when using stratified random sampling to estimate accuracy and area of land-cover change. *Remote Sensing Letters*, 3, 111–120.
- Stehman, S. V. (2013). Estimating area from an accuracy assessment error matrix. *Remote Sensing of Environment*, 132, 202–211.
- Stehman, S. V., & Czaplewski, R. L. (1998). Design and analysis for thematic map accuracy assessment: Fundamental principles. *Remote Sensing of Environment*, 64, 331–344.
- Stehman, S. V., & Foody, G. M. (2009). Accuracy assessment. In T. A. Warner, M.D. Nellis, & G. M. Foody (Eds.), *The SAGE handbook of remote sensing*. London: Sage Publications.
- Stehman, S. V., Olofsson, P., Woodcock, C. E., Herold, M., & Friedl, M.A. (2012). A global land cover validation dataset, II: Augmenting a stratified sampling design to estimate accuracy by region and land-cover class. *International Journal of Remote Sensing*, 33, 6975–6993.
- Stehman, S. V., & Selkowitz, D. J. (2010). A spatially stratified, multi-stage cluster sampling design for assessing accuracy of the Alaska (USA) National Land-Cover Data (NLCD). *International Journal of Remote Sensing*, 31, 1877–1896.
- Stehman, S. V., Sohl, T. L., & Loveland, T. R. (2003). Statistical sampling to characterize recent United States land-cover change. *Remote Sensing of Environment*, 86, 517–529.
- Stehman, S. V., & Wickham, J.D. (2011). Pixels, blocks of pixels, and polygons: Choosing a spatial unit for thematic accuracy assessment. *Remote Sensing of Environment*, 115, 3044–3055.
- Stehman, S. V., Wickham, J.D., Wade, T. G., & Smith, J. H. (2008). Designing a multi-objective, multi-support accuracy assessment of the 2001 National Land Cover Data (NLCD 2001) of the conterminous United States. *Photogrammetric Engineering & Remote Sensing*, 74, 1561–1571.
- Strahler, A. H., Boschetti, L., Foody, G. M., Friedl, M.A., Hansen, M. C., Herold, M., et al. (2006). Global land cover validation: Recommendations for evaluation and accuracy assessment of global land cover maps. *EUR 22156 EN – DG*. Luxembourg: Office for Official Publications of the European Communities (48 pp.).
- Tomppo, E. O., Gschwantner, T., Lawrence, M., & McRoberts, R. E. (2010). *National forest inventories: Pathways for common reporting*. New York: Springer.
- UN-REDD (2008). UN Collaborative Programme on Reducing Emissions from Deforestation and Forest Degradation in Developing Countries (UN-REDD). *FAO, UNDP, UNEP Framework Document*.
- Wickham, J.D., Stehman, S. V., Fry, J. A., Smith, J. H., & Homer, C. G. (2001). Thematic accuracy of the NLCD 2001 land cover for the conterminous United States. *Remote Sensing of Environment*, 114, 1286–1296.
- Wickham, J.D., Stehman, S. V., Gass, L., Dewitz, J., Fry, J. A., & Wade, T. G. (2013). Accuracy assessment of NLCD 2006 land cover and impervious surface. *Remote Sensing of Environment*, 130, 294–304.
- Woodcock, C. E., Allen, R., Anderson, M., Belward, A., Bindschadler, R., Cohen, W., et al. (2008). Free access to Landsat imagery. *Science*, 320, 1011.
- Wulder, M.A., Franklin, S., White, J. C., Linke, J., & Magnussen, S. (2006). An accuracy assessment framework for large-area land cover classification products derived from medium resolution satellite data. *International Journal of Remote Sensing*, 27, 663–683.
- Wulder, M.A., Masek, J. G., Cohen, W. B., Loveland, T. R., & Woodcock, C. E. (2012). Opening the archive: How free data has enabled the science and monitoring promise of Landsat. *Remote Sensing of Environment*, 122, 2–10.
- Wulder, M.A., White, J. C., Coops, N. C., & Butson, C. R. (2008). Multi-temporal analysis of high spatial resolution imagery for disturbance monitoring. *Remote Sensing of Environment*, 112, 2729–2740.
- Wulder, M.A., White, J. C., Hay, G. J., & Castilla, G. (2008). Towards automated segmentation of forest inventory polygons on high spatial resolution satellite imagery. *The Forestry Chronicle*, 84, 221–230.
- Wulder, M.A., White, J. C., Luther, J. E., Strickland, L. G., Rempel, T. K., & Mitchell, S. W. (2006). Use of vector polygons for the accuracy assessment of pixel-based land cover maps. *Canadian Journal of Remote Sensing*, 32, 268–279.
- Wulder, M.A., White, J. C., Magnussen, S., & McDonald, S. (2007). Validation of a large area land cover product using purpose-acquired airborne video. *Remote Sensing of Environment*, 106, 480–491.
- Zimmerman, P. L., Housman, I. W., Perry, C. H., Chastain, R. A., Webb, J. B., & Finco, M. V. (2013). An accuracy assessment of forest disturbance mapping in the western Great Lakes. *Remote Sensing of Environment*, 128, 176–185.