OBJECT-BASED CLOUD AND CLOUD SHADOW DETECTION IN LANDSAT IMAGES FOR TROPICAL FOREST MONITORING

Maciel Zortea, Arnt-Børre Salberg, Øivind Due Trier

Norwegian Computing Center
P.O. Box 114 Blindern
NO-0314 Oslo - Norway
{Maciel.Zortea, Arnt-Borre.Salberg, Oivind.Due.Trier}@nr.no

KEY WORDS: Cloud detection, shadows, classification, segmentation, Landsat

ABSTRACT:

Clouds and cloud shadows often obscure parts of images acquired by optical space-borne sensors. The clouds and cloud shadows need to be detected and labeled as missing data. This enables subsequent methods to make their own decisions about how the missing data should be handled. Here we propose an automatic method to detect daytime cloud and cloud shadows in the context of tropical forest monitoring. In particular, we focus on Landsat 5 TM and Landsat 7 ETM+ images. In addition to the original bands, we investigate the use of additional spectral-derived features, based on pixel-wise differences, ratios, and maximum values derived for all combinations of pairs of top-of-the atmosphere reflectance bands. The subset of features retained for classification, and the boundaries of the classes in the feature space, were identified by optimizing the accuracy of the proposed method using samples collected from spatially disjoint scenes, acquired in different time periods, in an attempt to increase the generalization capability of the proposed approach when applied to unseen scenes. When a new image is to be classified, the idea is to first segment it locally using the Statistical Region Merging algorithm (Nock and Nielsen, 2004). Cloud and cloud shadow masks are then obtained by classifying the averaged pixel values, inside each segment, instead of individual pixels. Finally a simple cloud shape matching algorithm is used to reduce false detection of cloud shadow areas. We found that the proposed object-based technique reduces the spatial noise of the final classified map when compared to traditional single pixel classification. The accuracy of the proposed method appears to be comparable to two alternative algorithms selected for benchmark purposes.

1 INTRODUCTION

For clouds, the top-of-atmosphere reflectance in the visible and near-infra red bands tend to be brighter than the surrounding background surface. This difference is more pronounced when the background land cover is vegetation. Clouds also tend to be colder. The literature of cloud detection on optical imagery is rich. Traditional cloud detection techniques classify the spectral pixel values individually, without using contextual information. Example of alternative strategies for cloud detection include the use of spectral clustering (Simpson and Gobat, 1996), or modeling of spatial information using Markov Random fields (Le Hégarat-Mascle and André, 2009), among many others.

Robust detection of cloud shadows is not trivial, but good estimates may be obtained by using information about the viewing geometry and the sun location (Le Hégarat-Mascle and André, 2009). Currently, the Norwegian Computing Center is developing methods for vegetation classification and change detection from time-series of optical images with missing data, e.g., due to cloud cover (Salberg, 2011a, Salberg and Trier, 2011). The detection of cloud and cloud shadows in an important component of the processing chain.

2 DATA AND METHODS

2.1 Landsat TM/ETM+ scenes

A set of 14 Landsat scenes acquired in Tanzania between 1987 and 2010 were used for developing and testing the method. The data was collected in 11 different path/rows, with 10 scenes acquired by Landsat 5 (TM) and 4 scenes by Landsat 7 (ETM+) (Tab. 1). The scenes were split into three disjoint sets, used for training, validation, and test purposes. The different geographic locations (path/row), in addition to distinct acquisition dates, allow us to better access the generalization capability of the proposed algorithm.

2.2 Protocol for labeling ground truth samples

Ground truth samples were manually selected by visual inspection of an RGB color composition of the Landsat scenes generated using the TOA reflectance bands {5,4,3}. Pixels corresponding to clouds, cloud shadows, and background (mostly vegetation, soil, and water) were selected using the polyline functionality available in region-of-interest tool in the ENVI software. In this laborious activity, typically the resulting number of labeled samples variates across the classes and scenes. In a final step, we randomly subsample the manually labeled regions selecting 500 pixels of each class, for each scene. This was done to give the same importance to all classes and scenes.

2.3 Development of the proposed algorithm: statistical classifiers and feature selection

The six visible and near-infrared bands of Landsat TM/ETM+ were first converted to top-of atmosphere reflectance (TOA). It is often the case that secondary features, that can be derived from the combination of the available bands, can also be useful for classification purposes. In the context of analysis of vegetation, a well know example that illustrates this idea is the use of the Normalized Difference Vegetation Index (NDVI). The NDVI is calculated as (NIR-VIS)/(NIR+VIS), where VIS and NIR stand for the spectral reflectance measurements acquired in the visible (red) and near-infrared bands.

In the case of cloud detection, ratios and/or differences of reflectance bands are sometimes employed (Irish et al., 2006). Given the six reflectance bands available in Landsat, we test three

Table 1: The Landsat images used in this study, and the estimated area of cloud and cloud shadow for the scenes [in %], according to the different algorithms tested: LEDAPS, the proposed OB-C (object based classification), and the GMM (Gaussian mixture model)

	#	Landsat	Path/Row	Year-Month-Day	Solar Zen.(Azi.)	LEDAPS	OB-C	GMM	$\cap_{\texttt{clouds}}$	OB-C	GMM	$\cap_{\texttt{shadows}}$
Training	1	5	168 / 62	1987-02-25	41 (101)	3.4	2.4	2.8	1.6	3.3	4.0	2.0
	2	5	172 / 63	2008-06-22	41 (46)	3.8	1.2	2.0	1.2	8.2	8.7	6.6
	3	5	165 / 68	1995-03-30	45 (73)	38.5	25.6	33.8	24.5	13.4	6.8	6.1
	4	7	167 / 65	2001-08-27	36 (61)	7.6	12.2	12.7	7.1	11.9	13.5	10.1
Validation	5	5	168 / 67	1994-10-10	36 (87)	45.8	20.0	37.7	19.1	1.6	3.1	1.0
	6	5	170 / 65	2010-11-05	27 (110)	0.0	0.6	6.5	0.0	0.0	0.3	0.0
	7	5	171 / 62	2010-01-28	36 (118)	32.7	2.1	9.4	1.9	9.2	6.5	2.9
	8	5	171 / 63	2007-10-03	27 (90)	35.0	38.5	35.2	30.8	14.5	14.0	12.4
Test	9	7	167 / 63	2001-03-04	33 (95)	2.2	0.7	1.4	0.4	0.8	0.6	0.5
	10	7	167 / 63	2001-05-07	36 (53)	13.3	11.2	12.6	9.8	12.9	11.4	10.2
	11	7	167 / 63	2001-12-01	32 (126)	10.5	15.1	21.6	9.6	5.8	4.7	3.6
	12	5	166 / 67	2009-07-17	45 (46)	17.5	16.8	18.3	14.8	11.4	12.2	10.0
	13	5	166 / 67	2009-06-15	45 (42)	24.2	24.0	24.5	21.4	13.2	12.6	11.0
	14	5	166 / 63	2009-11-06	28 (117)	12.9	11.8	11.2	7.6	48.1	8.6	7.2

typologies of features for possible use in the cloud and cloud shadow detection problem. We compute simple features relating pairs of reflectance bands, specifically the ratio $\{r(i,j) = \text{band } i/\text{band } j\}$, difference $\{\Delta(i,j) = \text{band } i - \text{band } j\}$, and maximum $\{\max(i,j) = \max(\text{band } i, \text{band } j)\}$, of all the possible combinations of bands, where (i,j) are the reflectance bands $\{1,2,3,4,5,7\}$. The above features are computed in a pixel-wise fashion, for all the 15 distinct pairs of bands that can be obtained. In addition to this large set of derived features, we use the 7 bands available, plus the NDVI, resulting in a set of $7+1+15\times 3 = 53$ features available for discrimination purposes.

Not all these features are equally relevant for the classification task. Obviously, the selection of a reduced subset would make computations faster, and reduce computer memory requirements. We would like the best subset of features to be (automatically) identified. The selected features might depend on the statistical classifier selected.

Feature selection was the approach adopted for feature reduction, and it was implemented using the Sequential Forward Selection algorithm (SFS) (Pudil et al., 1994). SFS relies on two key components (a) an objective function, called criterion, which is to be optimized, and (b) a search mechanism. Starting with an empty set, the SFS algorithm tests and adds sequentially one feature at a time to the candidate set until the addition of further features do not improve the criterion. SFS is intrinsically a suboptimal search algorithm, because an exhaustive search for all the possible combinations of features is computationally prohibitive. The average classification accuracy, measured in the validation set (spatially and temporally disjoint), is used as the criterion to guide the iterative search procedure.

Six different classifiers were tested in this study. The selection included both parametric and non-parametric approaches, which are among the classics in the pattern recognition literature (Hastie et al., 2009):

- Trees: decision tree classifier (CART algorithm)
- 1NN: the nearest neighbor approach using Euclidian distances in the feature space
- Naive Bayes: assume that features are independent of one another within each class. Here each feature is modeled using a univariate normal distribution
- Mahalanobis: uses the Mahalanobis distance with covariance estimates stratified by class
- LDA: linear discriminant analysis fits a multivariate normal density to each class, with a pooled estimate of covariance matrix
- QDA: quadratic discriminant analysis fits multivariate normal densities with covariance estimates stratified by class

The classifiers were trained with the training samples from the three classes of interest: (1) clouds, (2) background, (3) cloud shadows. During our computations, we assumed equal priors for all classes.

Selection of the best subset of features in a spatially and temporarily disjoint set is an attempt to increase the generalization of the classification model.

The accuracies and features, as selected by SFS, are shown in Tab. 2. Note that the statistical model used by each classifier is distinct, and this has an impact on the predicted classification accuracy and the ranked features. Results suggest that three features capture most of the accuracy that can be obtained. It is often the case that ratio and difference features were selected with higher priority than the original bands. For our current training and validation sets, the thermal band appears among the first 7 features selected when the parametric models are selected. This was not the case for the non-parametric classifiers tested.

The non-parametric nearest neighbor approach appears to perform slightly better than the other classifiers tested. This is not surprising if one considers the multimodal distributions of the training samples (Fig. 1), which are difficult to model with the Naive Bayes, LDA and QDA which are all unimodal. The performance of k-NN, for higher values of k is similar (not shown). At the current stage, we had to discard the option of using the nearest neighbor classifier due the lack of a fast implementation. Instead, we selected the Naive Bayes classifier, that despite its simplicity, gave better accuracies than its conceptually improved Mahalanobis, LDA, and QDA counterparts. The tree classifier would also be an option to consider.

2.4 Cloud and cloud shadow detection

For the purpose of this study, we focus on the Naive Bayes classifier, running only on the first three reflectance features identified in Tab. 2, namely band 5, band 1, and the difference between

Table 2: Sequence of first seven features selected by the SFS algorithm, and the respective average accuracy measure in the validation set used as the optimization criterion, for each of the classifiers trained using the training samples

Method	1	2	3	4	5	6	7
Trees	80.7 $[r(2,5)]$	95.8 $[B1]$	96.2 $[\Delta(1,2)]$	96.2 $[r(3,5)]$	96.5 $[r(4,7)]$	96.6 $[B3]$	96.6 $[B7]$
1NN	80.8 $[B2]$	96.1 $[\Delta(1,5)]$	97.0 $[\Delta(1,3)]$	97.2 $[B1]$	97.2 $[B7]$	97.2 $[\Delta(2,3)]$	97.3 $[\Delta(1,2)]$
NaiveBayes	81.7 $[B5]$	93.7 $[B1]$	96.1 $[\Delta(3,5)]$	96.2 $[\Delta(2,5)]$	96.4 $[\Delta(2,4)]$	96.7 $[B6]$	96.5 $[B2]$
Mahalanobis	68.0 $[r(2,5)]$	83.6 $[B2]$	86.3 $[r(4,5)]$	91.9 $[B4]$	91.3 $[\Delta(3,5)]$	88.5 $[B6]$	86.7 $[r(2,3)]$
LDA	74.7 $[B4]$	82.3 $[r(1,2)]$	85.1 $[r(1,5)]$	89.6 $[\Delta(2,5)]$	91.5 $[NDVI]$	92.5 $[r(3,5)]$	93.1 $[B6]$
QDA	81.7 $[B5]$	95.9 $[B1]$	93.7 $[B7]$	94.3 $[r(3,5)]$	94.4 $[B6]$	94.3 $[B3]$	94.8 $[r(1,3)]$

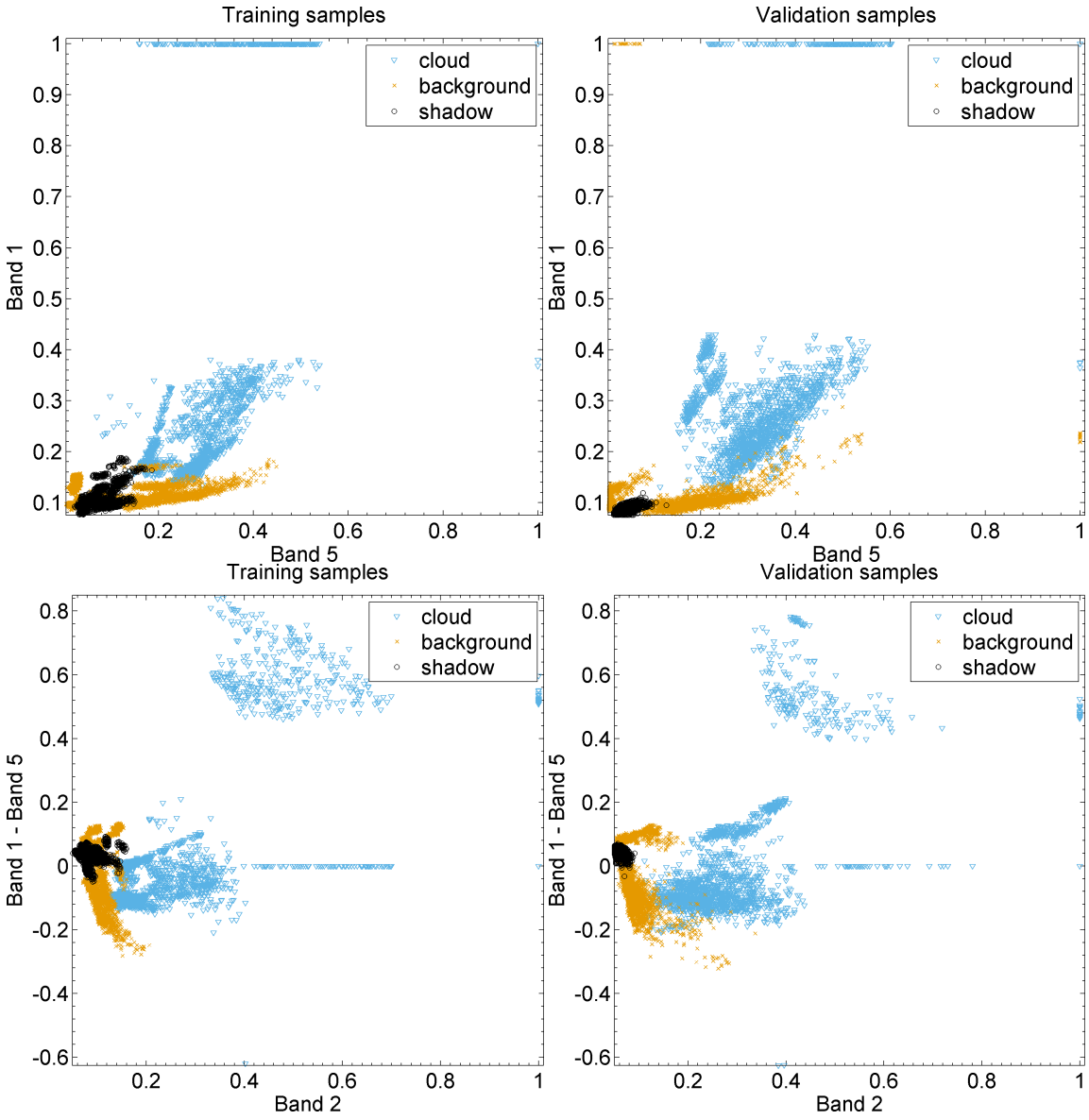


Figure 1: Training and validation samples shown in the first two features selected by the SFS algorithm using the Naive Bayes (top row) and 1–NN classifiers (bottom). The background class includes samples of soil, vegetation, water, among others. The training, validation, and test samples are collected in spatially disjoint scenes (see Tab. 1), with distinct temporal acquisitions. Notice the problem of shift of the class location in the feature space in the training and validation set, which is also the case for the test set (not shown), which makes the design of a robust classification algorithm challenging.

bands 3 and 5. The results in Tab. 2 suggest that this combination of bands provide good discrimination results.

At one hand, the use of only three features is convenient for generation of an RGB image that can be segmented using standard approaches that work with color images, like the Statistical Region Merging algorithm tested in this study (Nock and Nielsen, 2004). In our current proposal, this comes at the sacrifice of discarding the thermal band (and other features) that would potentially improve a bit the final classification. We hope to incorporate additional bands during the segmentation stage in future developments of the proposed methodology. By checking the typical range of reflectance values for the bands 5 and 1 (Fig. 1), it appears that saturation thresholds could be used to generate a RGB color composition without loss in the classification accuracy. We set the saturation thresholds at $[0.0, 0.6]$, $[0.0, 0.45]$, and $[-0.2, 0.2]$ for the first three selected features, respectively. We map linearly the resulting reflectance-based image using 8 bits per color channel. The color image is then segmented using the Statistical Region Merging (SRM) algorithm. SRM has an internal parameter Q that controls the coarseness of the segmentation. In our experiments we set $Q = 256$.

After the color image segmentation, for each segment, the average values for each spectral band are computed and used to classify each segment as cloud, shadow, or background. Notice that the classifier is trained with the original training samples, not with the segmented image. Our conjecture is that segment-based classification may be more robust than traditional pixel-wise classification if the averaged pixel values from the segments move away from the proximity of the class boundaries, alleviating a bit the problem of shift of the class location in the feature space in the test set.

Next, a simple algorithm matches the shapes of the clouds and the cloud shadows. This is done by translating the cloud mask on top of the segments classified as shadow until the best spatial shape match is found. Only the shadow segments that overlap (at least partially) the cloud mask are retained. Despite simple, this approach, that assumes approximately equal displacement of shadows with respect to the corresponding clouds, eliminates a lot of false cloud shadow detections, reducing substantially the spatial noise and false detections of cloud shadows.

2.5 Alternative algorithms for benchmark

We compare the cloud and cloud shadow detection results with an alternative approach based on a Gaussian mixture model (GMM) (Salberg, 2011b). The main idea is that the distributions of the test data can be obtained directly from each component of the mixture distribution after the corresponding parameters have been updated (from the distribution of the training data) using a low-rank dataset shift modeling scheme. Two reflectance bands, and the thermal band (which is resampled to 30m resolution) are used. After the parameters of the class distributions have been adapted to the test data, the image is classified using a Gaussian maximum likelihood classifier, regularized spatially by a Markov random field in order to obtain smooth class boundaries. The cloud shadow is identified by template matching in the sun azimuth direction. The resulting cloud and cloud shadow masks are also dilated to further remove any cloud/shadow remainings.

In addition, the proposed cloud algorithm is compared with the cloud mask produced by LEDAPS (Masek et al., 2006)[1]. The LEDAPS cloud mask include also a separate flag (QA). The missing values indicated in QA often appeared to be related to pixels with a saturated reflectance values. [2]

[1] the LEDAPS files labeled "lndcsm"

[2] We observed that this was likely to correspond to thick clouds. In order to make the comparison of the cloud masks as fair as possible, we checked each pixel with a missing value in QA. For those pixels, if both the proposed methods agree as a cloud pixel, and LEDAPS had a flag indicating an anomaly with the reflectance, then the LEDAPS cloud mask was updated to cloud too. Visual inspection of the "updated" cloud mask by LEDAPS suggested that our correction was realistic. A cloud shadow mask for LEDAPS is currently not available for comparison.

3 EXPERIMENTAL RESULTS

Table 3: Confusion matrix for the LEDAPS algorithm (in %). In this case the shadow mask is not available for comparison

	cloud	background	shadow
cloud	2859 (95.3)	141 (4.7)	-
background	6 (0.1)	4994 (99.9)	-
shadow	3 (0.1)	2997 (99.9)	-

Table 4: Confusion matrix for the GMM algorithm

	cloud	background	shadow
cloud	2941 (98.0)	59 (2.0)	(0.0)
background	25 (0.5)	4680 (93.6)	295 (5.9)
shadow	(0.0)	182 (6.1)	2818 (93.9)

Table 5: Confusion matrix for the OB-C algorithm (proposed)

	cloud	background	shadow
cloud	2739 (91.3)	261 (8.7)	(0.0)
background	9 (0.2)	4450 (89.0)	541 (10.8)
shadow	(0.0)	76 (2.5)	2924 (97.5)

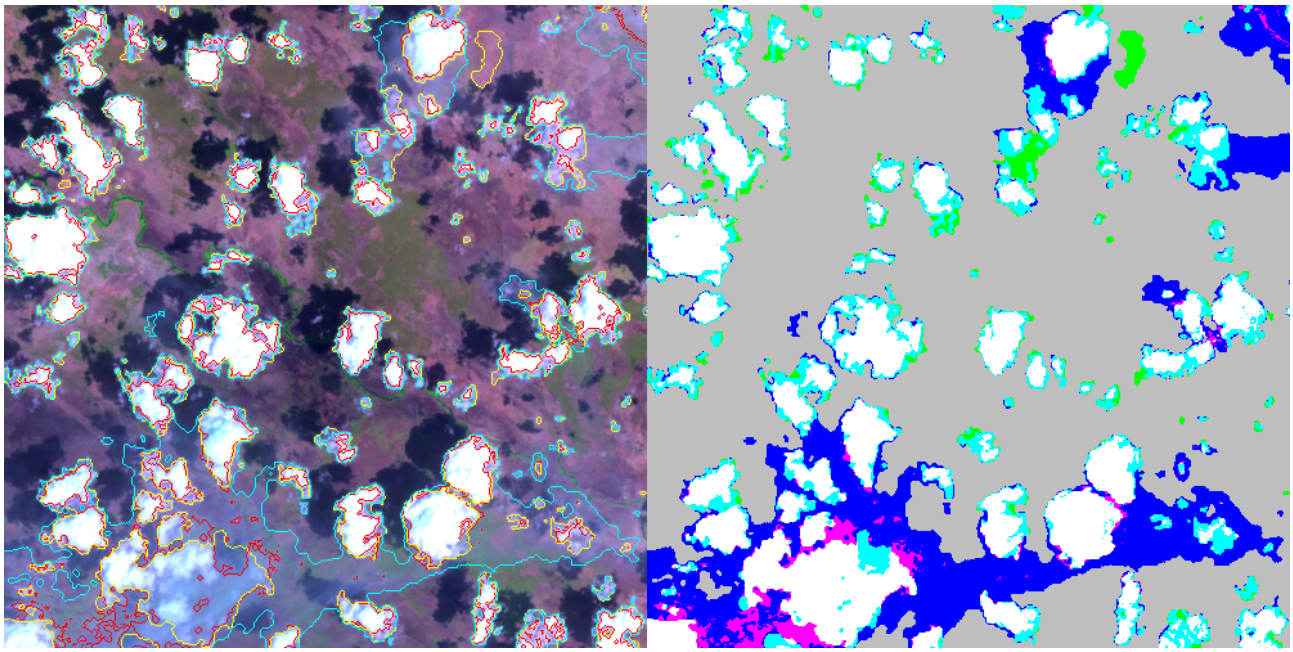
The accuracy for all the three cloud detection algorithms, measured in an independent test set, was found promising. The GMM algorithm performed best (98.0%), followed by LEDAPS (95.3%), and the proposed object-based method (91.3%) (Tabs. 3–5). In Tabs. 3–5, the entry (i, j) represents the count of test samples whose ground truth is the class i and whose predicted class is j. [3]

Conversely, when it comes to detection of cloud shadows, the accuracy of the proposed algorithm was found to be higher (97.5%) than the benchmark GMM algorithm (93.9%) (Tab. 4–5).

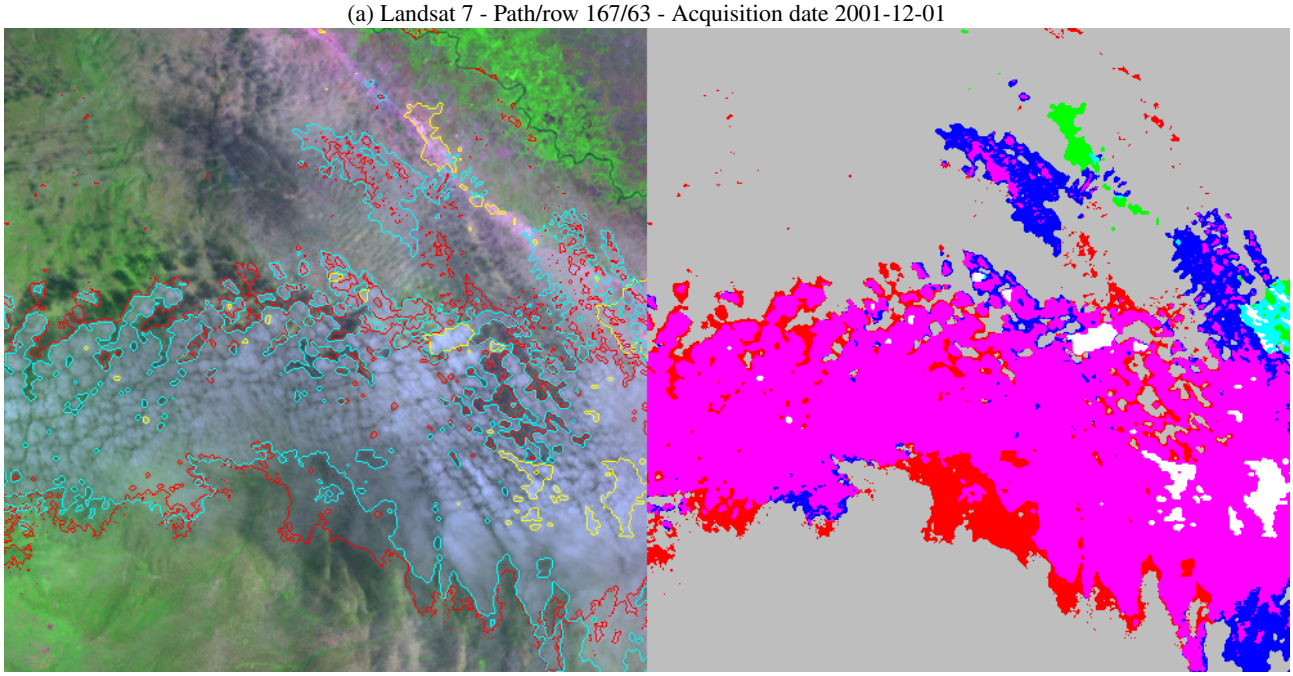
The three methods tested appear to detect for thick clouds well. Detection of haze remains very difficult, for all the algorithms tested, and the variability of the solutions in such conditions is typically very high (Fig. 2). The proposed OB-C appears to detect too little haze (Fig. 2). GMM provided reasonable detection of haze over distinct backgrounds (Fig. 2). We also found that LEDAPS can provide cloud masks that are spatially very noisy, especially on mountain regions. A careful inspection of the images revealed that the proposed algorithm confuses shadows with water. This was particulary the case for a large portion of sea confused with shadows (image #14 in Tab. 1). Confusion of shadow with water is also unavoidable for the alternative GMM approach.

Summary statistics for cloud area, and cloud shadow area are included in Tab. 1, with the respective agreements for all the scenes analyzed. The current figures were computed excluding a 50 pixel buffer zone around the borders of the images. In general the average values for each scene are in reasonable agreement, but visual inspection revealed that local agreement can be low in some of the scenes tested. It is important to stress that a careful

[3] The number of test samples for both the cloud and cloud shadow classes are 6 scenes x 500 = 3000 samples. In a similar way, 6 x 500 samples were collected for the background class. Additionally, water samples were identified in 4 test images, adding 4 x 500 = 2000 samples to the background class (5000 in total).

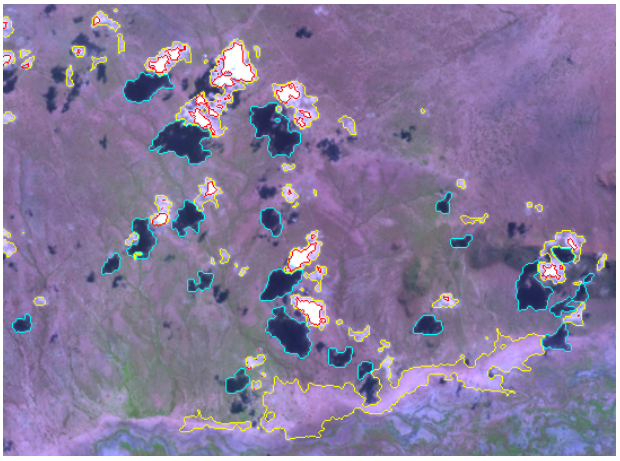


(a) Landsat 7 - Path/row 167/63 - Acquisition date 2001-12-01



(b) Landsat 7 - Path/row 167/063 - Acquisition date 2001-03-04

Figure 2: (Left) RGB composite (bands 5,4,3) showing the cloud boundaries, for LEDAPS (red), the proposed OB-C method (yellow), and the GMM approach (cyan). (Right) Map showing the agreement/disagreement of the cloud solution for these three methods. The white areas correspond to regions were all the three methods agree as cloud. The color key for interpretation of the other colors on the map is in Tab. 6.



Landsat 7 - Path/row 167/63 - Acquisition date 2001-12-01

Figure 3: RGB composite (bands 5,4,3) showing the cloud boundaries for LEDAPS (red) and the proposed OB-C method (yellow), as well as the detected shadows for the proposed OB-C method (cyan).

Table 6: Color key for the interpretation of the maps in Fig. 2

LEDAPS	Proposed OB-C	GMM	Color
–	–	–	gray
–	–	cloud	blue
–	cloud	–	green
–	cloud	cloud	cyan
cloud	–	–	red
cloud	–	cloud	magenta
cloud	cloud	–	yellow
cloud	cloud	cloud	white

visual inspection of the cloud and cloud mask results easily reveal several mistakes for all the methods tested. This suggests that robust cloud and cloud shadow classification is very challenging.

Our current, non-optimized, implementation of the proposed method takes about half an hour to run on a personal desktop (Intel Core i7 CPU at 3.4 GHz, 16 GB of RAM, 64-bit operating system). The color segmentation of the RGB derived Landsat scenes, which are typically about 8000 $\times$ 7000 pixels in size, using the Statistical Region Merging algorithm, is the most computationally intensive component of proposed methodology.

Additional examples of the cloud and cloud shadow results obtained using the proposed methodology are shown in Fig. 3. Some cases of "not very dark" shadows were missed, this happened also for the GMM method (not shown). In this particular case we observe that a large soil region was wrongly classified as cloud. The use of the thermal band in addition to the current features used for classification, or possibly the use of a non-parametric classifier, might improve the results.

4 DISCUSSION AND CONCLUSIONS

We have evaluated different candidate features and classifiers for for automatic cloud and cloud shadow detection with the ultimate goal of monitoring tropical forests in Tanzania. Our experimental results, focused on daytime Landsat TM/ETM+ scenes, revealed differences between the proposed object-based algorithm and the two alternative approaches tested, that can be used to further improve the proposed algorithm.

It is important to keep in mind that the accuracy scores shown in the confusion matrices are a simple attempt to rank the relative performance of algorithms tested, rather than provide an accurate estimate of the classification accuracy for all the scene. To obtain an unbiased estimate of the cloud and cloud detection accuracies when the proposed method would eventually become operational, it would be desirable to measure it using a true random sampling procedure for selection of test samples. Unfortunately this is not feasible at the time of writing this article.

Detection of haze/thin clouds remains very challenging for all the methods tested. The proposed object-based cloud and cloud shadow detection method seems to be detecting thick clouds and their associated cloud shadows well. However, the method needs improvement to be able to detect thin clouds and haze. The inclusion of the thermal band appears desirable.

One option to try to reduce the confusion between water and cloud shadows regions could be to model both classes separately, instead of grouping water with the background class. In some cases, maps with the location of the water bodies might be already available from other information sources, like a GIS layer.

Experiments using different path/row scenes would be desirable in order to extrapolate the general observations made here to different areas of tropical forests.

ACKNOWLEDGMENT

The authors thank USGS for providing Landsat images free of charge.

REFERENCES

Hastie, T., Tibshirani, R. and Friedman, J., 2009. *The Elements of Statistical Learning: data mining, inference, and prediction*. Springer Verlag: New York.

Irish, R., Barker, J., Goward, S. and Arvidson, T., 2006. Characterization of the Landsat-7 ETM+ automated cloud-cover assessment (ACCA) algorithm. Photogrammetric Engineering and Remote Sensing 72(10), pp. 1179.

Le Hégarat-Mascle, S. and André, C., 2009. Use of Markov random fields for automatic cloud/shadow detection on high resolution optical images. ISPRS Journal of Photogrammetry and Remote Sensing 64(4), pp. 351–366.

Masek, J., Vermote, E., Saleous, N., Wolfe, R., Hall, F., Huemmrich, K., Gao, F., Kutler, J. and Lim, T., 2006. A landsat surface reflectance dataset for North America, 1990-2000. IEEE Geoscience and Remote Sensing Letters 3(1), pp. 68–72.

Nock, R. and Nielsen, F., 2004. Statistical region merging. IEEE Transactions on Pattern Analysis and Machine Intelligence 26(11), pp. 1452–1458.

Pudil, P., Novovičová, J. and Kittler, J., 1994. Floating search methods in feature selection. Pattern Recognition Letters 15(11), pp. 1119–1125.

Salberg, A., 2011a. Land cover classification of cloud-contaminated multitemporal high-resolution images. IEEE Transactions on Geoscience and Remote Sensing 49(1), pp. 377–387.

Salberg, A., 2011b. Retraining maximum likelihood classifiers using a low-rank model. In: IEEE International Geoscience and Remote Sensing Symposium, pp. 166–169.

Salberg, A. and Trier, O., 2011. Temporal analysis of forest cover using hidden Markov models. In: IEEE International Geoscience and Remote Sensing Symposium, pp. 2322–2325.

Simpson, J. and Gobat, J., 1996. Improved cloud detection for daytime AVHRR scenes over land. Remote sensing of environment 55(1), pp. 21–49.