Cloud-Free Satellite Image Mosaics with Regression Trees and Histogram Matching

E.H. Helmer and B. Ruefenacht

Abstract

Cloud-free optical satellite imagery simplifies remote sensing, but land-cover phenology limits existing solutions to persistent cloudiness to compositing temporally resolute, spatially coarser imagery. Here, a new strategy for developing cloud-free imagery at finer resolution permits simple automatic change detection. The strategy uses regression trees to predict pixel values underneath clouds and cloud shadows in reference scenes from other scene dates. It then applies improved histogram matching to adjacent scenes. In the study area, the islands of Puerto Rico, Vieques, and Culebra, Landsat image mosaics resulting from this strategy permit accurate detection of land development with only spectral data and maximum likelihood classification. Between about 1991 and 2000, urban/built-up lands increased by 7.2 percent in Puerto Rico and 49 percent in Vieques and Culebra. The regression tree modeling and histogram matching require no manual interpretation. Consequently, they can support large volume processing to distribute cloud-free imagery for simple change detections with common classifiers.

Introduction

Persistent cloud cover over many regions complicates remote sensing with optical satellite imagery. Applications may require cloud-free parts from many scene dates for each vegetation map or for each of the times that bound a change detection. The variously dated scenes or cloud-free scene parts that might compose an image mosaic will differ in atmospheric conditions, sun-target-sensor geometry, sensor calibration, soil moisture, and vegetation phenology. These differences cause the relationships between land-cover classes and pixel brightness values to vary across space over a *mosaic period*, which refers to the time period spanning the cloud-free scenes or scene parts that compose an image mosaic.

One solution to variable relationships between land-cover classes and their spectral signatures is to separately classify scene dates (Achard and Estreguil, 1995; Cohen *et al.*, 2001; Helmer *et al.*, 2002). Depending on the number of scenes and their degree of cloudiness, however, a scene wise approach to vegetation mapping or change detection may not be practical. For example, where cloud-free imagery is common, scene footprints yield a regular arrangement of between-scene differences across space. This regular arrangement leads to a regular arrangement of the unique combinations of scene dates through time. Consequently, change detection can occur piecewise across space (Cohen *et al.*, 2002). Piecewise change detection is analogous to scene wise image classification as a solution to between-date radiometric differences across a project area. Insofar as cloud cover determines an irregular arrangement of scene dates across space for each mosaic period, unique combinations of scene dates through time form a complex patchwork. Piecewise change detection becomes infeasible where clouds are persistent. Reducing across space these radiometric and phenological scene differences could permit change detection with one seamless image mosaic for each temporal endpoint. If the land-cover changes of interest are spectrally non-subtle, the image mosaics that bound change detection intervals may not require radiometric normalization to each other (Cohen *et al.*, 1998; Song *et al.*, 2001).

Other alternatives exist to mosaicing cloud-free parts of scenes, but they have limitations. The spatial distribution or optical depth of clouds, along with the spatial complexity of land cover, limit how well geostatistical interpolation can predict cloud-obscured pixel values or land cover (Rossi *et al.*, 1994). Microwave satellite imagery, which clouds do not obscure, provides an alternative to optical imagery. Yet, land-cover discrimination with microwave imagery can benefit from optical data (Rignot *et al.*, 1997). A method to create cloud-free optical image mosaics expands the options available for satellite-based remote sensing.

Approaches for cloud removal include satellite image compositing, which selects pixels for an output, composite image that are least likely to have cloud cover from among scenes acquired over a *compositing period* (Gatlin *et al.*, 1984; Holben, 1986). A common compositing criterion is to select pixels for the output image with the largest values of the normalized difference vegetation index [NDVI = (Near Infrared − Red)/(Near Infrared + Red)]. A drawback of image compositing is residual cloud contamination, but excellent methods exist that detect or correct for cloud contamination in composite images (Gutman *et al.*, 1994; Cihlar *et al.*, 1996), or haze in Landsat scenes (Zhang *et al.*, 2002). For imagery with high temporal resolution, Cihlar *et al.* (1996), for example, detect cloud-contaminated pixels in composites with four thresholds. These thresholds include (a) the maximum red band reflectance that is present in the data set from snow and ice-free land under clear sky, (b) positive and negative deviations from the expected, median NDVI for a

E.H. Helmer is with the International Institute of Tropical Forestry, USDA Forest Service, 1201 Calle Ceiba, Jardín Botánico Sur, Río Piedras, PR 00926-1113 (ehelmer@fs.fed.us).

B. Ruefenacht is with Red Castle Resources, Inc., USDA Forest Service-Remote Sensing Applications Center, 2200 West, 2300 South, Salt Lake City, UT 84119-2020 (bruefenacht@fs.fed.us).

Photogrammetric Engineering & Remote Sensing
Vol. 71, No. 9, September 2005, pp. 1079–1089.

0099-1112/05/7109–1079/$3.00/0

pixel over the compositing period, and (c) a maximum deviation below the expected maximum NDVI for a pixel over a compositing period. Designed for image composites, this approach to cloud screening is not immediately applicable to imagery with low temporal resolution where only a few dates of cloud-free data may be available for a pixel. Nor does it explicitly address cloud shadow. A related issue is that dependence on image composites can limit spatial resolution. Compositing usually occurs over short time periods of 5 to 32 days to minimize phenological differences between the input scenes. The daily image acquisition that consistent phenology requires for a composite is only widely available for imagery with coarser spatial resolution. Image compositing itself degrades spatial resolution because of minor misregistrations between pixels from different scene dates. Misregistrations have potential to permeate an image more thoroughly in image composites than in mosaics of scene parts, because image composites are like pixel-level mosaics.

Zhang *et al.* (2002) developed an approach that optimally transforms Landsat images to detect the spatial distribution and intensity of haze and thin clouds. The transform quantifies the perpendicular displacement of a pixel from a *clear line*. This clear line forms in two-dimensional spectral space from the spectral signatures, in TM bands 1 and 3, of clear-sky pixels that span the range of land-cover classes present in a scene. Thick clouds or cloud shadows, however, will still obscure the scenes.

The objective of this study is to test whether a new strategy for developing cloud-free imagery over a project area can yield image mosaics that permit simple change detection. Core parts of the strategy require no manual interpretation and can thereby contribute to automatically processing large image volumes. For each scene in a project area and for each mosaicing period, this new strategy first uses regression trees to predict the image digital numbers (DNs) underneath clouds and cloud shadows in reference Landsat scenes from cloud-free parts of other scenes. The second part of the strategy is to mosaic the adjacent cloud-free scenes with histogram matching based on image overlap areas. Comparing this strategy for developing cloud-free imagery with other approaches will be important for evaluating its efficacy. However, this study focuses on a realistic application, which is another important aspect of testing a strategy. After developing cloud-free mosaics for each of two mosaic periods, we test whether the mosaics are suitable for simple change detection. Change detection with two image mosaics that span multiple scene footprints avoids the need for piecewise change detection. The change detection maps land-cover change to urban/built-up land, a process which we hereinafter also refer to as *land development*, using only spectral data and a maximum likelihood classifier.

Background

Where thick cloud cover persists, or when applications require finer spatial resolution, optical imagery, mosaicing cloud-free parts of all scene dates available may be the best option for cloud-free coverage. Techniques that might reduce radiometric differences between scene dates in an image mosaic include atmospheric correction, relative radiometric normalization, and histogram matching. Atmospheric correction converts image digital numbers (DNs) to absolute reflectance at Earth's surface (Chavez, 1996; Kaufman *et al.*, 1997). It first converts the DNs for each band to at-satellite radiance with sensor gains and offsets. It secondly converts at-satellite radiance to at-surface reflectance through correcting for atmospheric and solar effects.

Radiometric normalization calibrates images to each other with linear regression (Schott *et al.*, 1988; Vogelmann, 1988; Hall *et al.*, 1991; Olsson, 1993; Oetter *et al.*, 2001; Song *et al.*, 2001; Du *et al.*, 2002). For each band, a pixel-level model generally calibrates one or more *subject* scenes to a *reference* scene. The model has the general form in Equation 1:

$$y_{refi} = f(x_{subji}) \tag{1}$$

In Equation 1, y_{refi} and x_{subji} are brightness values for the i^{th} band from pixels in the reference and subject scenes, respectively, and they are usually co-located. The function $f(x_{subji})$ is most commonly linear, but the set of pixels used in the model varies. Equation 2 then estimates each radiometrically normalized pixel, y_{matchi}:

$$y_{matchi} = f(x_{subji}). \tag{2}$$

Most previous work on radiometric normalization focuses on normalizing the scene dates that bound change detections in a way that preserves phenological differences and avoids changing scenes to the point where normalization obscures change detection. Consequently, previous normalization models exclude pixels with the marked spectral changes that might occur with changes in land cover or phenology, use only same-season imagery across space, and use linear models. This limitation excludes alternate-season image data that might be crucial to a single time of cloud-free coverage over an area.

Histogram matching determines a lookup table for each image band that causes its histogram to resemble that of a reference image. Basic histogram matching identifies an output DN for each input DN through equating histogram cumulative distribution functions (CDFs). If $g(y)$ is the CDF for the histogram of a reference image, and $f(x)$ is the CDF of the histogram to be matched to the reference, then the histogram matching function for each DN is $g^{-1}(f(x))$. If the histograms have unequal pixel numbers, multiplying the ratio of the total pixel number in the reference image to that in the subject image scales the histogram matching function (Richards, 1993). This scaling may negatively affect histogram matching. Extensive areas of very bright or dark pixels can also cause poor histogram matches. Previous work has avoided such problems by using manually-selected pixels from scene overlap areas to match the midpoints of subject scene histograms to the midpoint of the reference scene histogram (Homer *et al.*, 1997).

A drawback of all these radiometric-matching techniques is that they require scenes across space with consistent vegetation phenology and soil moisture. Between-date differences in vegetation or soil phenology will cause land-cover classes to have more variable spectral signatures. The relationships between corresponding spectral bands in differently dated scenes will vary by land cover or vegetation type and cause nonlinearities in Equations 1 and 2. Such nonlinear relationships will degrade how closely linear normalization can radiometrically match imagery. When original scenes differ phenologically, mosaics of scenes matched from existing approaches may not be amenable to simple automated processing algorithms.

Given the limitations that compositing, thick cloud cover, atmospheric correction and linear normalization impose, this study addresses the need for a strategy to develop cloud-free mosaics with finer resolution imagery that is less dependent on same-season imagery. The goal is to generate satellite image mosaics that are amenable to simple automatic classification approaches. The mosicing strategy does not mechanistically address the between-date differences in Landsat scenes. Instead, it empirically minimizes those differences using regression trees and histogram matching. The first part of this new strategy uses regression trees to reduce radiometric differences between cloud-free scene parts from different dates. Because regression trees can model complex nonlinear relationships, they should provide more flexibility for matching models and the image data they use. Regression trees use training data to form regression models at the terminal nodes of decision trees. Decision trees recursively partition data into smaller groups based on tests at tree nodes (Breiman *et al.*, 1984; Hansen *et al.*, 1996; Friedl and Brodley, 1997). Both are now common in remote sensing for mapping variables like vegetation type, tree canopy cover, forest structural attributes or impervious surface cover (Lawrence and Wright, 2001; Hansen *et al.*, 2002; Moisen and Frescino, 2002; Yang *et al.*, 2003). Regression trees have the potential, then, to partition the relationships between spectral bands in differently dated scenes into sets of relationships for each band. That capability permits different matching relationships for different spectral ranges. Assuming that different spectral ranges correspond to different land-cover classes or vegetation types, regression tree models could more accurately predict image data than existing radiometric normalization or matching approaches.

For each Landsat path/row of a project area, the regression tree procedure assumes that additional predictor, or subject, scene dates exist that are cloudless where the reference scene has clouds as well as cloudless in some areas where the reference scene is also cloudless. Regression trees model the relationships between each band in the reference scene and bands in each predictor, or subject scene. Preliminary work indicated smaller matching model error with multiple predictor bands, which Olsson *et al.* (1993) also found was true for linear normalization models. Consequently, the regression tree models for Landsat Thematic Mapper (TM) or Enhanced Thematic Mapper (ETM+) scenes have the following general form:

$$y_{refi} = f(x_{subj1}, x_{subj2}, x_{subj3}, x_{subj4}, x_{subj5}, x_{subj7}). \tag{3}$$

In Equation 1, y_{refi} is the DN of a pixel in the reference scene for the i^{th} band to be predicted, x_{subj1} is the DN of the TM or ETM+ band 1 of the corresponding pixel in the subject scene, x_{subj2} is the DN for band 2 of the subject, and so on. For areas where two predictor scene dates are cloudless, A and B for example, a regression tree model can use pixels from two subject scenes to predict each band in the reference scene, in which case models have the following general form:

$$y_{refi} = f(x_{subjA1}, x_{subjA2}, x_{subjA3}, x_{subjA4}, x_{subjA5}, x_{subjA7}, x_{subjB1}, x_{subjB2}, x_{subjB3}, x_{subjB4}, x_{subjB5}, x_{subjB7}). \tag{4}$$

The second part of the strategy is to match the histograms of adjacent scenes that have undergone cloud removal. Although regression tree models based on image overlap areas might successfully match adjacent scenes, preliminary work showed some detail loss from image areas that regression tree models predicted. In contrast, histogram matching caused little detail loss. Histogram matching is unlikely to closely match scenes with markedly different phenologies. Conceivably, however, if the reference scenes that undergo cloud removal have similar phenology, matching the histograms of imagery based on those scenes may be successful. An important difference exists between basic histogram matching and the approach in this strategy. Here, the histogram matching function derives only from image overlap areas. Using only overlap areas eliminates scaling errors that result from matching histograms with unequal pixel numbers, and it ensures that the histograms for determining the match cover similar terrain. Unlike previous work, however, the approach is entirely automatic.

Methodology

Study Area

Puerto Rico and two of its outer islands, Culebra and Vieques (Figure 1), are Caribbean islands that have perpetual cloud cover. Puerto Rico has experienced rapid urban expansion, which threatens forest conservation in its relatively unprotected lowland moist and dry ecological zones (Helmer, 2004). Although small (884 km^2), these islands require four scenes for Landsat coverage because they occur at the intersection of two Worldwide Reference System (WRS) paths and rows. With elevations ranging from sea level to 1325 m, the islands have complex vegetation. Forest formations range from subtropical dry and moist forests to wet and rain forests including cloud forests (Ewel and Whitmore, 1973). Consequently, image processing and analysis for island-wide remote sensing applications is challenging (Helmer *et al.*, 2002).

Cloud Elimination from Landsat Scenes

As described in more detail below, the data-mining program Cubist (**www.rulequest.com**) developed regression tree models that predicted pixel values in cloudy parts of reference scenes from co-located pixels in subject, predictor scenes. The scenes for the 2000 mosaic period were ETM+ and the scenes for the 1991 mosaic period were TM. The dates of the reference scenes included (a) 24 December 1991 and 27 March 2000 for WRS path/row 4/47–48, and (b) 19 August 1992, and 13 November 2000 for WRS path/row 5/47–48. All the other available scenes that centered on a given time aided cloud removal for that mosaic period. The dates of these images were 21 January 1985, 15 January 1986, 22 January 1988, 22 March 1988, 03 February 1991, 03 September 1991, 17 September 1999, 14 May 2000, 02 August 2000, 18 August 2000, 10 September 2000, 09 January 2001, 25 January 2001, 26 February 2001, 05 March 2001, 11 July 2001, 20 July 2001, and 27 July

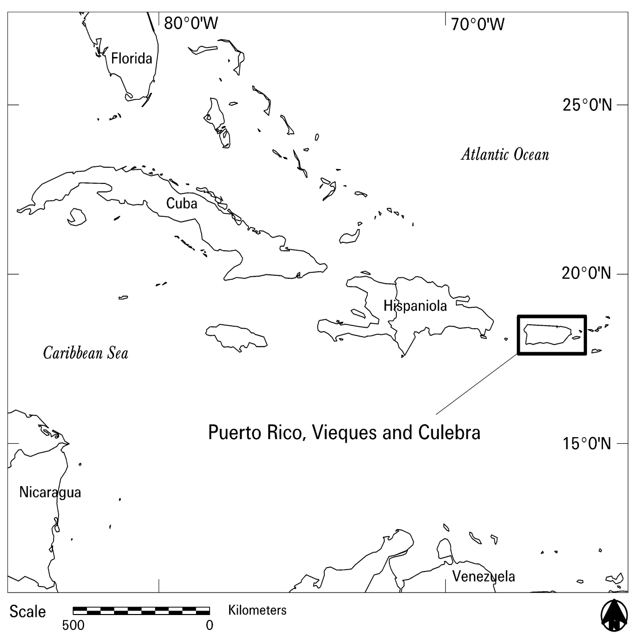

Figure 1. Study area location. The study area includes the Caribbean islands of Puerto Rico, Vieques and Culebra.

2001. The images had undergone terrain-corrected georeferencing that corrected parallax error from local topographic relief with a digital elevation model (Level 1T, **http://landsat7.usgs.gov/productinfo.html**). The program Image Tie Points (Kennedy and Cohen, 2003) generated control points for the scenes in each Landsat path/row to precisely co-register them. This method was accurate; the root mean square error between images ranged from 0 to 0.29 pixels. Imagery from 1985 and 1986 is six to seven years older than the target change detection endpoint date of 1991. However, most cloudy areas that these image dates predicted were at high elevations and included much reserve land. Because the area of land development in such locations is relatively small (Helmer, 2004), error from using these older scenes was also likely to be small.

Building Regression Tree Image Prediction Models

For each scene footprint and mosaic period, the cloud removal process began by making cloud, cloud shadow, and ocean water masks (cloud/shadow masks) for all scene dates. The cloud/shadow masks for this study derived from a combination of unsupervised classification with ISODATA, using ERDAS Imagine® v. 8.5 (Leica Geosystems, 2003) and manual editing. The second step was selecting a reference scene. Of those scenes for a mosaic period that were close to the target year for analysis, it was the least cloudy scene. Temporally close reference scenes over the study area, like same-dated scenes from a single Landsat path, or scenes with similar vegetation phenology, were preferable. Third, a second scene date served as the first subject scene for the footprint. It was *cloud complimentary* to the reference scene, meaning that it was cloud-free where the reference scene was cloudy.

The union of the cloud/shadow masks from the reference and subject scenes then masked both scenes, which revealed where both scenes were cloud-free. These mutually cloud-free areas supplied data for building the regression tree models in the form of six new images. Each new image contained the dependent and independent variables for a regression tree model that predicted one reference scene band. For example, the image for the model to predict band 1 in the reference scene had the following seven bands: y_{ref1}, x_{subj1}, x_{subj2}, x_{subj3}, x_{subj4}, x_{subj5}, and x_{subj7}, where the notation is the same as for Equation 3. Exporting each of these images to an ASCII file permitted their formatting for regression tree software. When two dates of subject scenes, A and B, for example, had common areas that were not obscured by clouds but were cloudy in the reference scene, the preceding steps incorporated a second subject image and its associated cloud/shadow mask. In that case, the six images for developing regression tree models each had 13 bands, which corresponded to the dependent variable and 12 independent variables in Equation 4.

Applying Regression Tree Models

The first step in applying the regression tree models was masking the subject scene with both its cloud/shadow masks and the inverse of the cloud/shadow mask for the reference scene. This step revealed where the reference scene was cloudy but the subject scene was clear. In these areas, the six regression models predicted six new DNs for pixels from corresponding pixels in the subject scene. Exporting and formatting the resulting image for input to regression tree models was the second step in applying the models. Integrating public domain code (**www.rulequest.com**) into an Imagine® program, with the Imagine® C Developer's Toolkit (Leica-Geosystems, 2003), enabled Imagine to interpret and apply the six regression tree models that predicted new DNs for each pixel. The predicted image parts then replaced areas in the reference scene that were cloudy. These steps were repeated for subsequent subject scenes that were clear where cloudy areas remained in the reference scene.

Summarizing Overall Errors for Each Scene that Underwent Cloud Removal

An independent error analysis estimated the combined mean and range of differences between reference image pixels and corresponding pixels that all regression tree models predicted. The observations for this analysis were from ten randomly selected, 1000-pixel areas that were cloudless in all image dates. These areas had been excluded from the pixels that went into the regression tree models. Each of the Cubist regression tree models that predicted data for a given scene predicted DNs for the 1000-pixel areas. That step permitted (a) finding absolute differences between reference image and predicted DNs, and (b) combining these differences to estimate mean potential errors by band and reference scene.

Histogram Matching with Image Match

The histogram matching used a new histogram matching technique, Image Match, within Imagine. The program matched images that had undergone cloud removal with images from adjacent scene footprints using one of those images from each mosaic period as a reference.

Image Match is a spatial model that automatically runs a series of image processing steps. It is identical to histogram matching based on equating cumulative distribution functions. However, it uses only image overlap areas to determine a lookup table for matching. Because these overlapping areas have equal total pixel numbers, the histogram match requires no scaling based on differences in total pixel numbers, and the terrain that each histogram covers is similar. For each band in two images to be matched, Image Match first finds where the two images overlap based on where they are both non-zero. It outputs data from each overlapping pixel to form two new images that are simply the overlapping parts of each input image. It then performs a basic histogram match between the overlapping areas using the RASTERMATCH function in Imagine, which outputs a matched image for the overlapping piece. The ZONAL_MIN raster function in Imagine then determines correspondence between DNs in the matched piece and the unmatched piece of the image to be matched. The function outputs a table of the minimum value in the matched piece for each of the 256 DNs in the unmatched piece. Because only one DN occurs in the matched piece for every DN in the unmatched one, zonal maximum or mean functions would give the same result. The resulting table is a lookup table that directly relates DNs in the original image to be matched to DNs in the matched overlapping piece. The DIRECT_LOOKUP function in Imagine® then outputs an image in which it replaces each DN in the original, entire image to be matched with the corresponding DN from this lookup table.

Change Detection with Hybrid Classification of Multitemporal Imagery

A simple hybrid supervised-unsupervised approach to detect land development began by merging optical bands and indices from both mosaic dates to form a multitemporal image (Howarth and Boasson, 1983; Nelson, 1983; Muchoney and Haack, 1994; Cohen and Fiorella, 1998). The definition of land development for the change detection was change to urban/built-up lands, mined lands, or bulldozed lands under preparation for development. Land development included any areas both within and outside of metropolitan areas if they changed from a vegetated surface to an impervious or bulldozed surface. It excluded lands that were bulldozed in the earlier date and urban/built-up by the later date. In addition to Landsat bands 1 through 5 and 7 for each mosaic period, we added NDVI and the band 4:5 ratio from each mosaic. The two indices are sensitive to vegetation; the band 4:5 ratio, for example, is sensitive to forest successional stage in both temperate (Fiorella and Ripple, 1993) and tropical (Helmer *et al.*, 2000) regions. Drops in their values are good indicators of land development. The resulting 16-band multitemporal image included six bands and two indices for each mosaic period. Second, an unsupervised ISODATA clustering (Leica Geosystems, 2003) classified the multitemporal image into a 25-class thematic image. The process also output a signature file containing the spectral signatures for each class.

Strategically displaying and enhancing three bands from the 16-band image revealed locations of land development (Howarth and Boasson, 1983; Sader and Winne, 1992; Cohen *et al.*, 1998; Seto and Liu, 2003). To visualize land development, we displayed TM band 1 from the first mosaic period in red, ETM+ band 1 from the second mosaic period in green, and NDVI from the earlier mosaic period in blue. Urban/built-up lands in both dates appeared yellow, and land development (i.e. change) appeared magenta. Agricultural fields that changed from mature or growing crops to bare soil were also magenta. Vegetated areas that didn't change were blue. Displaying band 1 in the red and green bands provided the most contrast between land development and other classes.

Displaying the clustered image simultaneously with the multitemporal image revealed two output classes that included land development. These two classes, however, were confused with (a) scattered pixels from urban/built-up land present in both times, for the first class, and (b) agricultural change, chiefly from mature sugar cane to bare cropland, as well as a few pastures in the driest areas that senesced and greatly brightened, for the second class. We replaced the signatures for these two confused classes with three new signatures, each extracted from 1,500 to 6,000 pixels, for (a) urban/built-up land present in both times,

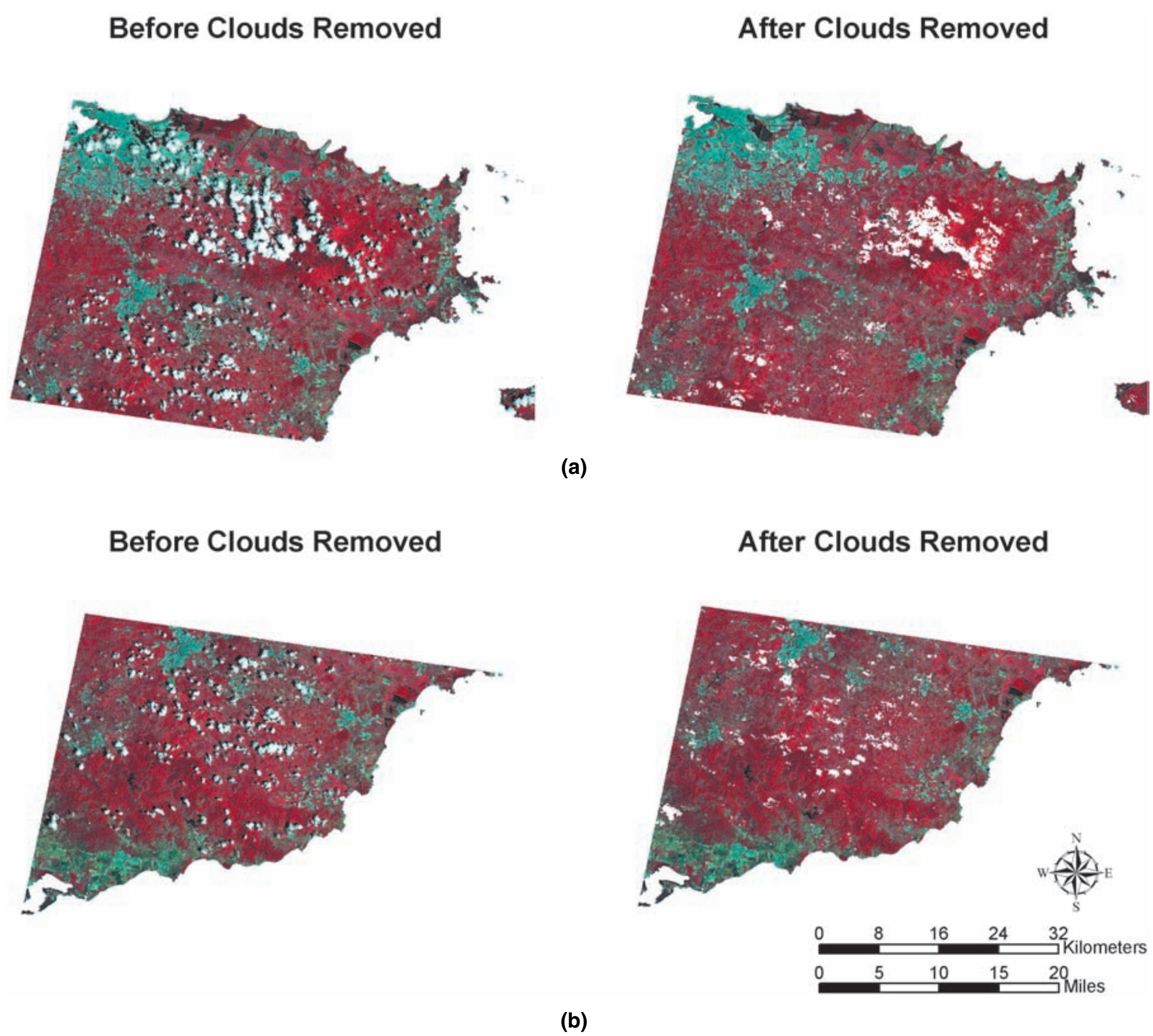

Plate 1. Results of cloud removal for eastern Puerto Rico for the image mosaic from 27 March 2000. Extensive cloud cover prevented filling all cloudy areas with data: (a) 27 March 2000 Landsat ETM+ Path 4, Row 47 (4,3,2) – 87 percent clouds removed, (b) 27 March 2000 Landsat ETM+ Path 4, Row 48 (4,3,2) – 83 percent clouds removed.

Table 1. Each of the Regression Tree Models that Predicted Cloudy Parts of a Given Reference Scene was also used to Predict Band Values in a Randomly-selected Set of 1,000 Cloud-free Pixels from that Scene. Band Wise Differences Between Predicted and Co-located Reference Scene Digital Numbers (DNs) Indicate How Closely the Regression Tree Models Predicted Reference Scene Data. Presented Below are the Mean and 95 percent Confidence Intervals for These Differences, Which are Averaged Over All the Models that Contributed to Predicting Cloudy Areas in a Given Reference Scene

Errors in Cloud Removal Process for Puerto Rico (Mean Difference between Reference and Regression tree Predicted DNs)

Landsat Path/Row	Landsat Thematic Mapper Band Number 1	2	3	4	5	7
2000 Mosaic period						
005/047	4 ± 0.5	5 ± 0.5	6 ± 0.7	9 ± 0.8	10 ± 0.9	7 ± 0.7
005/048	4 ± 0.4	5 ± 0.5	7 ± 0.7	8 ± 0.7	10 ± 0.9	7 ± 0.7
004/047	7 ± 0.9	8 ± 0.9	12 ± 1.3	9 ± 0.8	15 ± 1.5	12 ± 1.2
004/048	4 ± 0.8	4 ± 0.9	14 ± 1.2	10 ± 0.6	18 ± 1.6	14 ± 1.2
1991 Mosaic period[1]						
005/047–048	4 ± 0.4	2 ± 0.2	4 ± 0.4	10 ± 0.8	10 ± 0.8	5 ± 0.4
004/047–048	3 ± 0.3	2 ± 0.2	3 ± 0.3	7 ± 0.6	7 ± 0.6	4 ± 0.4

[1]Scenes for the 1991 mosaic period had been ordered as movable scenes centered on Landsat Rows 47 and 48.

(b) land development, and (c) the few areas of pasture with dramatic phenological change. The edited, 26-signature file enabled a supervised maximum likelihood classification of the 16-band multitemporal image. Manual editing removed confusion between land development and agricultural change in crop development stage from the resulting, 26-class map. After recoding all classes but land development to zero and performing a contiguity analysis, we removed all pixel clusters that had fewer than 11 pixels, yielding a 0.99 ha-minimum mapping unit (MMU). Only 2.4 percent of patches were smaller than this MMU, and this was a patch size that, based on fieldwork, was identifiable with certainty in accuracy assessment. Another reason for the MMU was that pixels that were a mixture of urban/built-up and undeveloped lands were likely to cause some misclassification of smaller patches. In Puerto Rico, such mixed pixels are easily misclassified in a maximum likelihood classification of TM spectral data alone (Helmer *et al.*, 2002).

Accuracy Assessment

A randomly selected set of 200 points included 100 points each from (a) lands that underwent land development between about 1991 and 2000, including change to bulldozed or mined land, and (b) lands that did not undergo land development between the two times. Aerial photos from both times permitted us to label each point as land development or no land development. The aerial photos included 1:32 000 scale NASA Aerochrome IR photos dated 1991 and 1:48 000 scale color photos that NOAA collected in 1999. With the goal of ensuring that we could correctly identify land development in aerial photos and imagery, fieldwork 09 April 2003 circumnavigated about three quarters of Puerto Rico along major roads or highways and selected side roads. It relied on integrating a GPS receiver with a laptop computer (with a daylight-viewable image display) running the ERDAS Imagine® GPS tool (Leica Geosystems, 2003).

Results

For the four scenes that cover Puerto Rico, the cloud removal procedure removed 73 to 87 percent of clouds and cloud shadows for the year 2000 mosaic period (Plates 1 and 2). Cloudy areas in the reference scenes were filled with new data from other scene dates (Plate 3). For the 1991 mosaic period, the procedure removed 18 percent of clouds in western Puerto Rico and 92 percent of clouds in the east. For some areas, none of the available images were cloud-free, which prevented the procedure from removing all clouds. The mean absolute differences between reference pixel values and the values that the various regression models predicted for each scene ranged from 2 to 18 DNs, but most were ≤10 DNs (Table 1). These differences estimate overall errors in the regression tree procedure for each scene and band.

Matching the histograms of adjacent images that underwent cloud removal worked well. The new adjacent images matched tonally to the base image. Mosaicing adjacent images produced a virtually seamless mosaic, which facilitated change detection. Visually comparing image mosaics, with and without prior histogram matching with Image Match (Plate 4), showed that images mosaiced without the procedure had visible seam-lines.

The hybrid unsupervised-supervised change detection that manually recoded agricultural change (Plate 5) was accurate (Table 2), correctly classifying 85.4 percent of points and yielding an error matrix with a Kappa coefficient of agreement of 0.66 ± 0.12. A remarkable 49 percent increase in patches of urban/built-up lands ≥1 ha on the two small outer islands provided a first and vital estimate of land development there over the decade (Table 3).

Discussion

Cloud-free Image Mosaic Strategy

Regression trees have successfully modeled continuous forest- and land-cover attributes from satellite imagery (Hansen *et al.*, 2002; Moisen and Frescino, 2002; Yang *et al.*, 2003). In this study they estimate new image DNs. Using regression trees to model the relationships between co-located pixels from different image dates is new both in its use of regression tree models and its application of such predictive calibration models across space.

Previous work (Homer *et al.*, 1997) uses data from scene overlap areas to match adjacent Landsat scene histograms. The strategy in this study uniquely emphasizes an automatic way to apply histogram matches from image overlap areas. Deriving a lookup table for histogram matching from the histograms of overlapping areas reliably produces better histogram matches that are not subject to scaling errors. Scaling adjustments are otherwise requisite for matching histograms with unequal pixel numbers.

Table 2. The Two Image Mosaics that Resulted from Applying a New Strategy to Make Cloud-free Mosaics were Merged into One Multitemporal Image. A Simple Hybrid Supervised-Unsupervised Classification of the Multitemporal Data Identified Change to Urban/Built-up lands (Land Development). Shown Below is the Accuracy of the Change Detection, Which is Based on Aerial Photo Interpretation of 200 Randomly Selected Change and No-change Points (N = 192 after Eliminating Points that Were Cloudy in Aerial Photos)

Percent Correct Overall	Kappa Coefficient of Agreement	Commission Error for Mapped Land Development	Omission Error for Mapped Land Development
85.4	0.66 ± 0.12	15.5	12.0

Table 3. Areas of Urban/Developed Land-Cover in about 1991 and 2000, and Increase in Urban/Developed Patches ≥1 ha from 1991 to 2000. The Change Detection Permitted the First Estimates of Land Development for the Study Area that are Based Solely on Landsat Imagery. The Large, 49 Percent Increase in Urban/Developed Lands in Vieques and Culebra is the First Estimate of its Kind for those Islands. Total Area for the Islands is About 874,120 ha for Puerto Rico and 15,385 ha for Vieques and Culebra

	Puerto Rico	Vieques, Culebra	Total
Urban/developed 1991 (ha)	91,799[1]	180	97,379
Urban/developed 1991–2000 (ha)	6,647	89	6,736
Total in 2000 (ha)	98,446	269	98,715
Percent increase 1991–2000	7.2%	49%	7.3%

[1]Helmer *et al.*, 2002.

Empirically rather than mechanistically minimizing the differences between Landsat scenes means that these differences are sources of residual error in the two-part mosaicing strategy. These between-date error sources include differences in season (which affect vegetation and soil phenology), and differences in solar azimuths, solar elevation angles, and atmospheric conditions. Additional error sources are non-uniform atmospheric conditions across a given scene and land-cover changes that occur during each mosaic period. Differences in illumination between scene dates may explain some of the detail loss that occurs with regression tree prediction. For example, topographic features can become less pronounced. Illumination differences could cause some pixels to be sunlit in one date and shadowed in another, which could cause regression tree models to brighten shadowed dark pixels and darken sunlit bright ones. Neither the regression tree models nor the histogram matches explicitly address within-scene differences in atmospheric conditions. Any non-uniform atmospheric conditions within scenes likely degrade how accurately the regression tree models can predict pixel values. Error in regression tree models may also arise from non-uniform changes in the spectral signatures of common land-cover classes. Such changes could arise from land-cover and agricultural changes during each mosaic period.

Cloud-free imagery resulting from this strategy will likely support detection of other spectrally non-subtle changes, like forest clearing or burning, assuming that time intervals are short enough to prevent confusion between regrowing and undisturbed forest. Accordingly, wide availability of such imagery could lead to more timely forest monitoring with the most basic image processing tools and skills. Yet, how the strategy compares with other possible strategies or performs with other classifiers or more complex classification objectives remains in question. For example, errors in the strategy may cause confusion between land-cover or change classes with subtle spectral differences. The degree to which more sophisticated classifiers can accommodate these errors, or altogether eliminate the need for this sort of strategy, remains unknown.

Many other aspects of the strategy merit further exploration or improvement. The regression tree prediction and histogram matching require no manual image interpretation. However, large-volume image processing will require fully automated methods to make cloud/shadow masks. Also somewhat unclear is whether the strategy would benefit from atmospherically correcting imagery prior to applying it. Such correction, though, should facilitate automatic cloud detection. In addition, incorporating an approach like that in Zhang *et al.* (2002), which permits correction for intra-scene haze variations, might decrease errors in the matching process. Another potentially important issue is the detail loss in image data output from regression trees. If illumination differences between scene dates contribute to this detail loss, perhaps incorporating related ancillary data into the regression tree models could alleviate this problem. Finally, only parts of the mosaicing strategy may be applicable to imagery with finer or coarser spatial resolution. The histogram matching may, for example, be more appropriate for images with fine spatial resolution if detail loss prevents the regression tree procedure from being useful.

Change Detection

Applying the two-part strategy to develop image mosaics makes possible a fast and accurate detection of land development with only spectral data and a maximum likelihood classifier. The resulting map quantifies land development between about 1991 and 2000 (Table 3). It further shows how land development has a spatial pattern that both extends and intensifies (Plate 5). A visual analysis of land development locations indicates that Puerto Rico remains similar to other temperate and tropical landscapes. In Puerto Rico, the strongest spatial predictors of land development are proximity to existing urban areas and roads. Higher elevations or slopes decrease land development likelihood. Nevertheless, topography loses importance for undeveloped lands remaining in metropolitan areas, where land development pressures are largest (Helmer, 2004). An outcome of these geographical drivers is an intensification pattern in which remaining undeveloped patches in and near urban or residential areas undergo development. Other studies have observed this pattern in Puerto Rico and elsewhere (Lugo, 2002; Yang *et al.*, 2003).

Although the change detection manually discriminates change from mature crops to bare soil, maximum likelihood classifiers can confuse these changes even when single scene dates, as opposed to mosaics of many scene dates, bound change detection intervals (Seto and Liu, 2003). Agricultural change is now limited in Puerto Rico's landscape, but manual editing may be impractical where agricultural spatial patterns are complex. In such landscapes, neural network classifiers that accommodate spectrally heterogeneous classes may distinguish agricultural change from land development (Seto and Liu, 2003). Additional times of imagery might also distinguish agricultural change if later images changed back from bare soil to mature crops.

Land-cover changes that occur during each mosaic period are an error source in the change detection, causing over- or

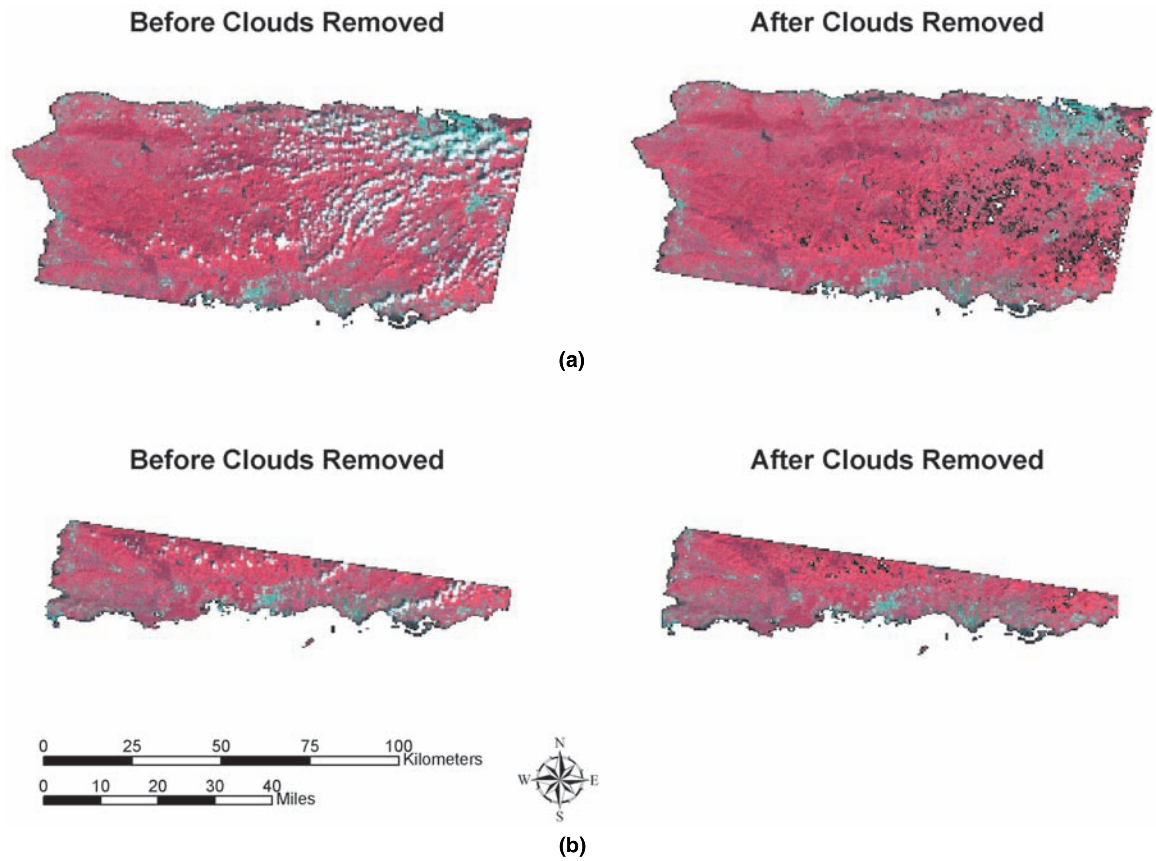

Plate 2. Results of cloud removal for western Puerto Rico for the image mosaic from about 2000. Extensive cloud cover prevented filling all cloudy areas with data: (a) 13 November 2000 Landsat EMT+ Path 5, Row 47–73 percent clouds removed, (b) 13 November 2000 Landsat EMT+ Path 5, Row 48–75 Percent clouds removed.

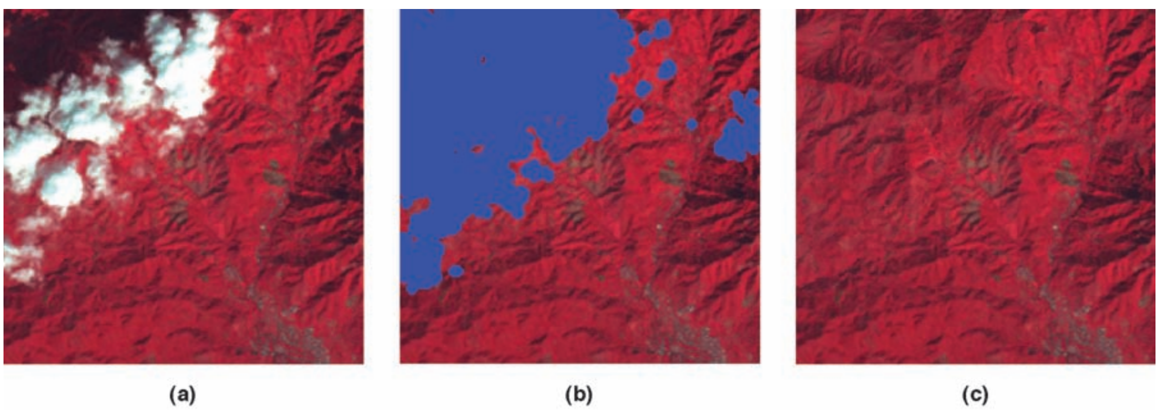

Plate 3. Clouds were removed from (a), the image with clouds, using regression tree models. The result is shown in image (c) in which the clouds were removed. Image (b) shows the areas (blue) where the regression tree models were applied, i.e., the image with its cloud/shadow mask.

under-estimates of change areas. Moreover, where clouds occurred in all scene dates available, land-cover change remains unknown. Despite the many scenes that formed each mosaic, and even though the data set included scenes from different seasons, some locations were very persistently cloudy. A second error source in the change detection is its 30-m spatial resolution. At this scale, pixels that contain a mixture of developed and vegetated lands cause error in

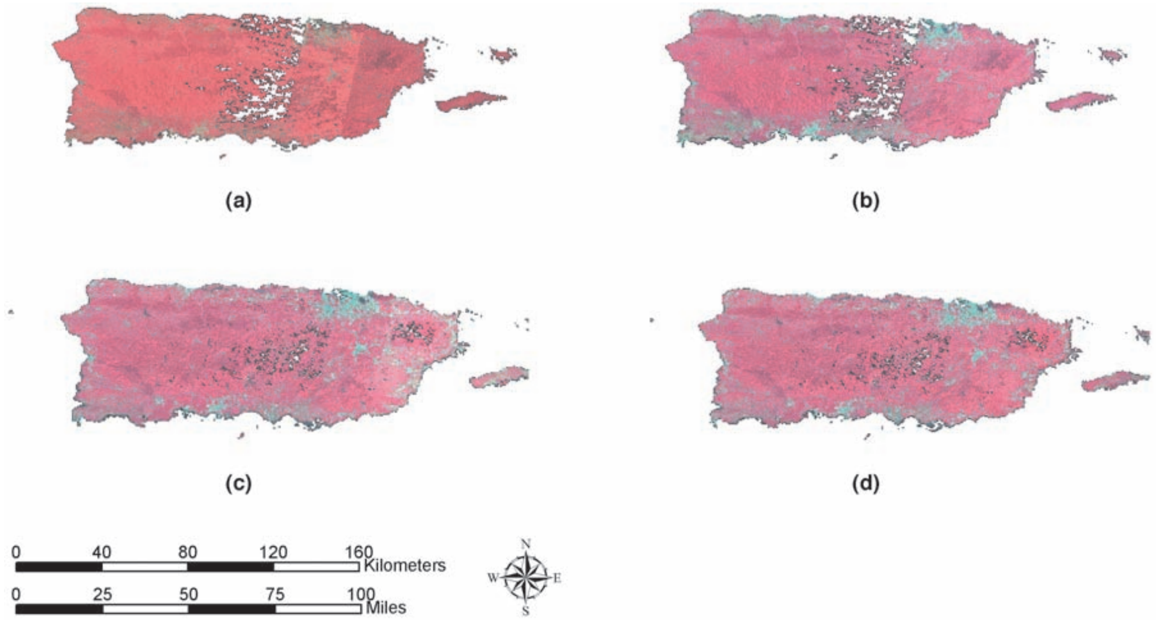

Plate 4. Results of mosaicing Landsat scenes of Puerto Rico compared with mosaicing scenes after using the Image Match histogram matching procedure. The images (a) 1991 mosaiced, and (c) 2000 mosaiced were only mosaiced, and they have visible seam-lines. The images (b) 1991 Image Matched, and (d) 2000 Image Matched were mosaiced with Image Match and have no visible seam-lines. The white areas are the remaining clouds/shadows.

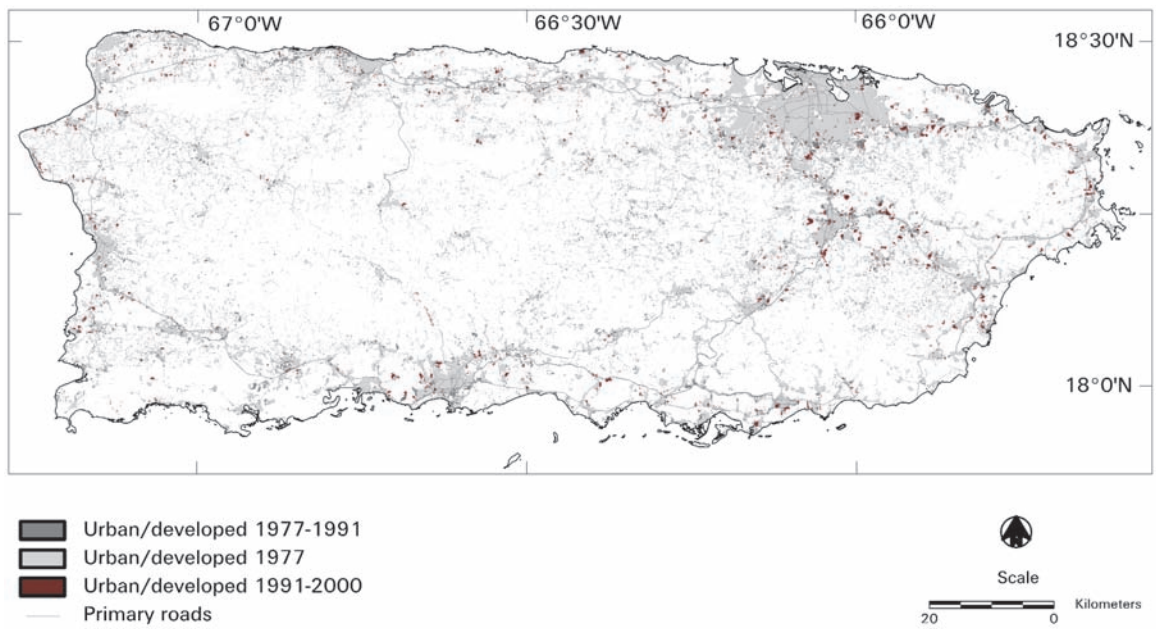

Plate 5. Land development in Puerto Rico from about 1991 to 2000 (this study) and about 1977 to 1991 (Helmer, 2004), and urban/developed areas in 1977 (Ramos and Lugo, 1994). Note that although accuracy assessment points did not cover Vieques and Culebra (not shown), we carefully visually verified the results with aerial photos and finer resolution satellite imagery.

land-cover classifications (Helmer *et al.*, 2002). The MMU that eliminated land development patches ≤1 ha avoided attempting to classify such pixels. The consequence is that the data are unlikely to include land development at the scale of single homes, which does occur along rural roads.

Summary and Conclusions

This paper presents a new strategy for making nearly cloud-free image mosaics with Landsat satellite imagery. The strategy first uses regression tree models to predict band values of cloudy pixels in a reference scene from other scene dates. It secondly matches adjacent scenes with histogram matching based only on image overlap areas. Results of the study indicate that regression tree prediction offers an effective tool for overcoming persistent cloud cover in Landsat imagery. In addition, histogram matching based on image overlap areas permits seamless mosaicing of scenes that have undergone cloud removal with regression tree prediction. Finally, this study shows that mosaics resulting from the new strategy can support change detection in persistently cloudy regions. A fast and accurate detection of change to urban/built-up lands, with only spectral data from two mosaics and a maximum likelihood classifier, demonstrates this conclusion. Errors in the regression tree predictions have the potential to increase confusion between classes with subtle spectral differences, and they cause some detail loss in imagery. Consequently, applying the strategy to form cloud-free image mosaics for complex classification objectives may require more sophisticated classifiers.

The regression tree modeling and histogram matching require no manual interpretation. Consequently, they can support large volume processing to distribute cloud-free imagery. Such imagery should permit simple change detections of spectrally marked changes, such as land development or forest clearing, with widely available classification tools.

Acknowledgments

The USFS Forest Inventory and Analysis Program and National Forest System funded the RSAC portion of this work. Landsat ETM+ imagery was provided through NASA (Martha Maiden, Woody Turner), USGS EDC (Larry Tieszen), and USFS IITF (E. Helmer, B. Gould). We thank Carmen Santiago, U.S. Natural Resources Conservation Service (Puerto Rico) for making aerial photos available from 1991, and Todd Kennaway and Brent Read of Colorado State University's CEMML, who manually recoded confused agricultural change. Dave Vanderzanden developed the initial spatial models for Image Match. Justin Gray configured information systems for fieldwork. We thank three anonymous reviewers for their valuable input. This research was conducted with cooperation from the University of Puerto Rico.

References

Achard, F., and C. Estreguil, 1995. Forest classification of Southeast Asia using NOAA AVHRR data, *Remote Sensing of Environment*, 54(3):198–208.

Breiman, L., J.A. Friedman, R.A. Olshen, and C.J. Stone, 1984. *Classification and Regression Trees*, Wadsworth International Group, Belmont, California, 358 p.

Chavez, P.S., 1996. Image-based atmospheric corrections – revised and improved, *Photogrammetric Engineering & Remote Sensing*, 62(9):1025–1036.

Cihlar, J., H. Ly, and Q. Xiao, 1996. Land cover classification with AVHRR multichannel composites in northern environments, *Remote Sensing of Environment*, 58(1):36–51.

Cohen, W.B., and M. Fiorella, 1998. Comparison of methods for detecting conifer forest change with Thematic Mapper imagery, *Remote Sensing Change Detection, Environmental Monitoring Methods and Applications* (C.D. Elvidge, Editor), Ann Arbor Press, Ann Arbor, Michigan, pp. 89–102.

Cohen, W.B., M. Fiorella, J. Gray, K. Anderson, and E.H. Helmer, 1998. An efficient and accurate method for mapping forest harvest activity in the Pacific Northwest, *Photogrammetric Engineering & Remote Sensing*, 64(4):293–300.

Cohen, W.B., T.K. Maiersperger, T.A. Spies, and D.R. Oetter, 2001. Modelling forest cover attributes as continuous variables in a regional context with Thematic Mapper data, *International Journal of Remote Sensing*, 22(12):2279–2310.

Cohen, W.B., T.A. Spies, R.J. Alig, D.R. Oetter, T.K. Maiersperger, and M. Fiorella, 2002. Characterizing 23 years (1972–1995) of stand replacement disturbance in western Oregon forests with Landsat imagery, *Ecosystems*, 5(2):122–137.

Du, Y., P.M. Teillet, and J. Cihlar, 2002. Radiometric normalization of multitemporal high-resolution satellite images with quality control for land cover change detection, *Remote Sensing of Environment*, 82(1):123–134.

Ewel, J.J., and J.L. Whitmore, 1973. *The Ecological Life Zones of Puerto Rico and the U.S. Virgin Islands*, *ITF-18*, Institute of Tropical Forestry, Rio Piedras, Puerto Rico, 72 p.

Fiorella, M., and W.J. Ripple, 1993. Determining successional stage of temperate coniferous forests with Landsat satellite data, *Photogrammetric Engineering & Remote Sensing*, 59(2):239–246.

Friedl, M.A., and C.E. Brodley, 1997. Decision tree classification of land cover from remotely sensed data, *Remote Sensing of Environment*, 61(3):399–409.

Gatlin, J.A., R.J. Sullivan, and C.J. Tucker, 1984. Considerations of and improvements to large-scale vegetation monitoring, *IEEE Transactions in Geosciences and Remote Sensing*, GE-22:496–502.

Gutman, G.G., A.M. Ignatov, and S. Olson, 1994. Towards better quality of AVHRR composite images over land: Reduction of cloud contamination, *Remote Sensing of Environment*, 50(2):134–148.

Hall, F.G., D.E. Strebel, J.E. Nickeson, and S.J. Goetz, 1991. Radiometric rectification: Toward a common radiometric response among multidate, multisensor images, *Remote Sensing of Environment*, 35(1):11–27.

Hansen, M., R. Dubayah, and R. Defries, 1996. Classification trees: an alternative to traditional land cover classifiers, *International Journal of Remote Sensing*, 17(5):1075–1081.

Hansen, M.C., R.S. DeFries, J.R.G. Townshend, R. Sohlberg, M. Carroll, and C. DiMiceli, 2002. Towards an operational MODIS continuous field of percent tree cover algorithm: examples using AVHRR and MODIS data, *Remote Sensing of Environment*, 83(1–2):303–319.

Helmer, E.H., 2004. Forest conservation and land development in Puerto Rico, *Landscape Ecology*, 19(1):29–40.

Helmer, E.H., W.B. Cohen, and S. Brown, 2000. Mapping montane tropical forest successional stage and land use with multi-date Landsat imagery, *International Journal of Remote Sensing*, 21(11):2163–2183.

Helmer, E.H., O. Ramos, T.d.M. Lopez, M. Quiñones, and W. Diaz, 2002. Mapping forest type and land cover of Puerto Rico, a component of the Caribbean biodiversity hotspot, *Caribbean Journal of Science*, 38(3–4):165–183.

Holben, B.N., 1986. Characteristics of maximum-value composite images from temporal AVHRR data, *International Journal of Remote Sensing*, 7:1417–1434.

Homer, C.G., R.D. Ramsey, T.C. Edwards, and A. Falconer, 1997. Landscape cover-type modeling using a multi-scene Thematic Mapper mosaic, *Photogrammetric Engineering & Remote Sensing*, 63(1):59–67.

Howarth, P.J., and E. Boasson, 1983. Landsat digital enhancements for change detection in urban environments, *Remote Sensing of Environment*, 13(2):149–160.

Kaufman, Y.J., A.E. Wald, L.A. Romer, B.C. Gao, R.R. Li, and L. Flynn, 1997. The MODIS 2.1 um channel – correlation with visible reflectance for use in remote sensing of aerosol, *IEEE Transactions on Geoscience and Remote Sensing*, 35:1–13.

Kennedy, R.E., and W.B. Cohen, 2003. Automated designation of tie-points for image-to-image coregistration, *International Journal of Remote Sensing*, 24(17):3467–3490.

Lawrence, R.L., and A. Wright, 2001. Rule-based classification systems using classification and regression tree (CART) analysis, *Photogrammetric Engineering & Remote Sensing*, 67(10): 1137–1142.

Leica Geosystems, 2003. *ERDAS Field Guide*, Leica Geosystems GIS & Mapping, LLC, Atlanta, Georgia, 688 p.

Lugo, A.E., 2002. Can we manage tropical landscapes? – an answer from the Caribbean perspective, *Landscape Ecology*, 17(7): 601–615.

Moisen, G.G., and T.S. Frescino, 2002. Comparing five modelling techniques for predicting forest characteristics, *Ecological Modelling*, 157(2–3):209–225.

Muchoney, D.M., and B.N. Haack, 1994. Change detection for monitoring forest defoliation, *Photogrammetric Engineering & Remote Sensing*, 60(10):1243–1251.

Nelson, R.F., 1983. Detecting forest canopy change due to insect activity using Landsat MSS, *Photogrammetric Engineering & Remote Sensing*, 49(9):1303–1314.

Oetter, D.R., W.B. Cohen, M. Berterretche, T.K. Maiersperger, and R.E. Kennedy, 2001. Land cover mapping in an agricultural setting using multiseasonal Thematic Mapper data, *Remote Sensing of Environment*, 76(2):139–155.

Olsson, H., 1993. Regression functions for multitemporal relative calibration of Thematic Mapper data over boreal forest, *Remote Sensing of Environment*, 46(1):89–102.

Ramos, O.M., and A.E. Lugo, 1994. Mapa de la vegetación de Puerto Rico, *Acta Científica*, 8(1–2):63–66.

Richards, J.A., 1993., *Remote Sensing Digital Image Analysis, An Introduction*, Springer-Verlag, New York, 340 p.

Rignot, E., W.A. Salas, and D.L. Skole, 1997. Mapping deforestation and secondary growth in Rondonia, Brazil, using imaging radar and thematic mapper data, *Remote Sensing of Environment*, 59(2):167–179.

Rossi, R.E., J.L. Dungan, and L.R. Beck, 1994. Kriging in the shadows: Geostatistical interpolation for remote sensing, *Remote Sensing of Environment*, 49(1):32–40.

Sader, S.A., and J.C. Winne, 1992. RGB-NDVI colour composites for visualizing forest change dynamics, *International Journal of Remote Sensing*, 13(16):3055–3067.

Schott, J.R., C. Salvaggio, and W.J. Volchok, 1988. Radiometric scene normalization using pseudoinvariant features, *Remote Sensing of Environment*, 26(1):1–14.

Seto, K., and W. Liu, 2003. Comparing ARTMAP Neural Network with the maximum-likelihood classifier for detecting urban change, *Photogrammetric Engineering & Remote Sensing*, 69(9):981–990.

Song, C., C.E. Woodcock, K.C. Seto, M.P. Lenney, and S.A. Macomber, 2001. Classification and change detection using Landsat TM Data: when and how to correct atmospheric effects?, *Remote Sensing of Environment*, 75(2):230–244.

Vogelmann, J.E., 1988. Detection of forest change in the Green Mountains of Vermont using Multispectral Scanner data, *International Journal of Remote Sensing*, 9(7):1187–1200.

Yang, L., G. Xian, J.M. Klaver, and D. Brian, 2003. Urban land-cover change detection through sub-pixel imperviousness mapping using remotely sensed data, *Photogrammetric Engineering & Remote Sensing*, 69(9):1003–1010.

Zhang, Y., B. Guindon, and J. Cihlar, 2002. An image transform to characterize and compensate for spatial variations in thin cloud contamination of Landsat images, *Remote Sensing of Environment*, 82(2–3):173–187.

(Received 29 January 2004; accepted 10 May 2004; revised 17 May 2004)