Estimating Percent Tree Cover Using Regression Tree Method with Very-High-Resolution QuickBird Images as Training Data

ROKHMATULOH*, Daisuke NITTO*, Hussam Al BILBISTI*, Kota ARIHARA* and Ryutaro TATEISHI*

Abstract

The estimation of tree cover area at continental scale is becoming more important than before due to the needs to improve our understanding of carbon dynamics. The estimation of percent tree cover of a large area using MODIS data by regression tree method is a promising method. New points of this study are the use of QuickBird images for the collection of training data and the use of the Stepwise Linear Regression (SLR) for selecting the best subset of predictor variables. The estimation of percent tree cover of African continent was tried using 11 QuickBird images to get 195 cells as training data and 32-day composite MODIS 2003 data as predictor variables. The predictor variables consist of surface reflectance, normalized difference vegetation index (NDVI), enhanced vegetation index (EVI), normalized difference soil index (NDSI) and land surface temperature (LST). The result shows that NDVI and surface reflectance bands are effective to estimate percent tree cover and this method is acceptable with the prediction error of 5.17%.

Keywords: tree cover, regression tree method, QuickBird, predictor variables

1. Introduction

The Kyoto Protocol sets a collective global target of reducing greenhouse gas emissions by about 5% of 1990 levels by the first commitment period of 2008 to 2012[1]. In order to enter into force, it needs to be ratified by at least 55 countries that are responsible for at least 55% of the countries’ carbon emissions. Carbon dioxide (CO₂) in the atmosphere is a greenhouse gas that contributes considerably to global warming[2]. One possible strategy to reduce greenhouse gases with great potential is to use trees to sequester CO₂[3]. The growing trees remove CO₂ from the atmosphere through the process of photosynthesis and store the carbon in plant structures[4]. Estimation of tree cover area at continental scale is becoming more important than before due to the needs to improve our understanding of carbon dynamics[5,6]. Different percentages of tree cover store different amounts of carbon and the changes in tree cover, as expressed in a greenness factor to effectively surrogate biomass, are used in the model to calculate the annual changes of carbon[7–9]. Percent tree cover is defined as the percentage of the ground surface area that is covered by a vertical projection of the outermost perimeter of the natural spread in the plants’ foliage; small openings in the crown are included[10].

Previous studies to estimate percent tree cover for global or continent basis have used spectral mixture analysis (SMA) and linear mixture model[11]. The SMA and linear mixture model techniques use a linear model to approximate the relationship between the spectral signal and canopy cover. However, such relationships are often very complex and highly variable, especially over large areas. A separate study of percent tree cover mapping was performed by scientists from the University of Maryland (UM) using 850 m MODIS data[12]. This study applied regression tree method, which was subsequently modified using a stepwise procedure and bias adjustment. However, for training data, they used Landsat MSS (80 m) and TM (30 m) data, where discrimination of tree and non-tree is relatively difficult. Another study used regression tree method and high-resolution of IKONOS as training data and analyzed for small areas[13]. Therefore, very-high-resolution image, such as QuickBird, for training data in a continent based mapping is required for better discrimination among tree, herbaceous, grass, urban and bare areas within mix-land cover areas[14,15].

Africa is a continent of diversity and forests constitute more than 17% of the world’s forests, covering 520 million hectares or almost 18% of the land area. Emissions of greenhouse gases that cause climate change are still low, estimated to be only 7% of global emission. In addition,
Estimating Percent Tree Cover Using Regression Tree Method with Very-High-Resolution QuickBird Images as Training Data

Africa’s vast forest reserves serve as a significant sink for CO₂ and thus play an important role in alleviating and balancing the emission of the industrialized countries[46]. However, recently forests in Africa are threatened by a combination of factors including agricultural expansion, heavy livestock grazing, and accelerated urbanization and industrialization[7].

This study describes efforts to develop a method to map percent tree cover of Africa with very-high-resolution of QuickBird images as training data or predicted variable. Explanatory or predictor variables were extracted from MODIS data. The percent tree cover information on an annual basis is an important parameter in order to implement the Kyoto Protocol as committed by the United Nations Framework Convention on Climate Change (UNFCC)[1]. The percent tree cover estimation was accomplished using a regression tree method. The regression-tree estimates a case’s target value in terms of its attribute values by constructing a model containing one or more rules, where each rule is a conjunction of conditions associated with a linear expression[18, 19]. An advantage of a regression tree is its ability to effectively use proportional or continuous predictor data sets with different measurement scales[19].

2. Data

A coarse resolution of MODIS/Terra Nadir BRDF-Adjusted Reflectance 16 Day L3 global 1 km SIN grid product (MOD43B4NBAR)[20] data (MODIS) with seven bands (band 1–7) were used as an input to cover a large area with daily acquisition. These MODIS land bands consist of red/band 1 (620–670 nm), near infrared (NIR)/band 2 (481–876 nm), blue/band 3 (459–479 nm), green/band 4 (545–565 nm), shortwave infrared (SWIR)/band 5 (1230–1250 nm), SWIR/band 6 (1628–1651 nm) and SWIR/band 7 (2105–2155 nm). MODIS data in 2003, starting from 3 December 2002 to 16 December 2003, were acquired from USGS in 10° x 10° of Sinusoidal format. These data were then mosaicked and reprojected into Geographic Latitude/Longitude to have the same projection as QuickBird data.

The 16-day MODIS bands data were composed to 32-day (monthly) composite data using maximum value composite (MVC) formula[21]. In addition to these monthly data, normalized difference vegetation index (NDVI), enhanced vegetation index (EVI)[22] and normalized difference soil index (NDSI)[23] were also extracted from MODIS data. The land surface temperature (LST)[24] 8-day composite 1 km data with the same period as MODIS 16-day were also included in explanatory or predictor variables for estimating percent tree cover. The LST data are required to geographically stratify between dry woodland and moist forest in Africa[12, 25, 26]. These LST data include two time observations, day and night data.

For training data, 11 images of QuickBird data were chosen to represent different land cover types according to Global Land Cover (GLC) 2000[17] map (Fig. 1). Ten images are distributed in Africa and one image in Asia to represent desert areas in the northern part of Africa; each image of QuickBird data covers an area of about 5 × 5 km. QuickBird data are satellite images that record high-resolution data in the visible and near-infrared range. The QuickBird data contain multispectral data with 2.44 m resolution and panchromatic data with 0.60 m resolution. In this study, the QuickBird pan-sharpened 0.60 m resolution data were used. An image of Landsat ETM (ETM) with 15 m resolution was also added to represent barren areas (land cover with zero percent tree cover) of Africa. An ETM image with 15 m resolution was produced using an image fusion technique between visible and panchromatic bands[28]. A list of QuickBird and ETM images employed in this study is described in Table 1.

3. Methodology

Modeling an empirical relationship between percent tree cover as a predicted variable and MODIS 1 km data as a predictor or explanatory variables was accomplished using regression tree methods. The models were then applied to all pixels in a mapping area to produce a per-pixel estimation of the percent tree cover over Africa. This method quantifies spatial distribution of tree cover as a continuous variable of 0 to 100%, and offers a repeatable technique to characterize percent tree cover for a global area.

A main approach in this study comprises four steps:

1. Deriving percent tree data from QuickBird images using an unsupervised k-means clustering or a supervised maximum likelihood classification[31] and a visual interpretation;
2. Constructing predictor variables such as surface reflectance, NDVI, EVI, NDSI, and LST derived from MODIS data;
3. Selecting the potentially most useful predictor variables based on Cp statistics[30] and finally,
4. Spatially extrapolating the developed models from Cubist tools[31] to the entire 1 km MODIS data of Africa using the National Land Cover Database (NLCD) mapping tool[32]. Fig. 2 illustrates a regression tree method used in this study.

Percent tree cover, which then will be used as ground truth data, from 11 QuickBird images were obtained firstly by an unsupervised clustering. If non-tree classes and shadowed
areas were difficult to be separated with tree class, then a manual editing and an on-screen digitizing were also performed to obtain better class-separation results. For cases in which the result of an unsupervised clustering was unsatisfactory e.g., the misclassified classes more than 10%, the tree and non-tree extraction was further modified by a supervised classification, followed by an on-screen digitizing (see Fig. 2). The predictor variables were extracted from the monthly-composited MODIS data that consist of surface reflectances of band 1 to band 7, NDVI, EVI, NDSI and LST during December 2002 to November 2003. NDVI is expressed as:

$$\text{NDVI} = \frac{(\text{NIR} - \text{red})}{(\text{NIR} + \text{red})} \quad (1)$$

where NIR is band 2 and red is band 1 of MODIS. EVI is expressed as:

$$\text{EVI} = \frac{\rho_{\text{NIR}} - \rho_{\text{red}}}{\rho_{\text{NIR}} + C_1 \rho_{\text{red}} - C_2 \rho_{\text{blue}} + L} \quad (2)$$

![Fig. 1 Location of QuickBird and ETM images as training data. The selection of QuickBird images was based on tree coverage, which was also considered different land cover types according to GLC 2000 map. The ETM image was added only to represent barren areas, which had zero-percent tree cover.](image)

**Table 1 QuickBird images employed in this study (11 images) and one additional ETM image.**

<table>
<thead>
<tr>
<th>Scene</th>
<th>Acquisition Date (yy/mm/dd)</th>
<th>Center Coordinate (Lat./Long.)</th>
<th>Off-Nadir Angle (Degrees)</th>
<th>GLC 2000 Land Cover Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2003/2/15</td>
<td>0.69/24.93</td>
<td>15</td>
<td>Tree cover, broadleaved, evergreen</td>
</tr>
<tr>
<td>2</td>
<td>2004/7/3</td>
<td>10.67/10.08</td>
<td>11</td>
<td>Mosaic: Cropland / tree cover /other natural vegetation</td>
</tr>
<tr>
<td>3</td>
<td>2003/12/7</td>
<td>6.69/16.89</td>
<td>6</td>
<td>Mosaic: Tree cover /other natural vegetation</td>
</tr>
<tr>
<td>4</td>
<td>2003/1/3</td>
<td>9.23/8.30</td>
<td>9</td>
<td>Mosaic: Cropland / shrub and/or grass cover</td>
</tr>
<tr>
<td>5</td>
<td>2003/1/16</td>
<td>14.47/-12.08</td>
<td>6</td>
<td>Shrub cover, closed-open, deciduous</td>
</tr>
<tr>
<td>6</td>
<td>2004/3/27</td>
<td>0.20/18.59</td>
<td>12</td>
<td>Tree cover, regularly flooded, fresh water</td>
</tr>
<tr>
<td>7</td>
<td>2007/12/5</td>
<td>-7.37/22.14</td>
<td>11</td>
<td>Tree cover, broadleaved, evergreen</td>
</tr>
<tr>
<td>8</td>
<td>2002/11/18</td>
<td>-29.99/27.53</td>
<td>5</td>
<td>Shrub cover, closed-open, deciduous</td>
</tr>
<tr>
<td>9</td>
<td>2002/6/28</td>
<td>-12.05/16.97</td>
<td>13</td>
<td>Tree cover, broadleaved, deciduous, open</td>
</tr>
<tr>
<td>10</td>
<td>2003/10/14</td>
<td>18.64/42.08</td>
<td>13</td>
<td>Sparse herbaceous or sparse shrub cover</td>
</tr>
<tr>
<td>11</td>
<td>2004/5/9</td>
<td>-0.33/37.56</td>
<td>19</td>
<td>Tree cover, broadleaved, evergreen</td>
</tr>
<tr>
<td>ETM</td>
<td>2003/3/19</td>
<td>31.91/30.31</td>
<td>-</td>
<td>Herbaceous (closed-open), bare, and urban</td>
</tr>
</tbody>
</table>
Estimating Percent Tree Cover Using Regression Tree Method with Very-High-Resolution QuickBird Images as Training Data

where, $\rho_{NIR}$ is NIR reflectance, $\rho_{red}$ is red reflectance, $\rho_{blue}$ is blue reflectance, $C_1$ is the atmosphere resistance red correction coefficient, $C_2$ is the atmosphere resistance blue correction coefficient, $L$ is the canopy background brightness correction factor and $G$ is the gain factor. The coefficients adopted in the EVI algorithm are, $C_1=6$, $C_2=7.5$, $L=1$, and $G=2.5^{33}$. The NDSI is expressed as $^{33}$:

$$NDSI = \frac{SWIR - NIR}{SWIR + NIR}$$

which SWIR is band 6 and NIR is band 2 of MODIS.

Whereas the regression tree allows a large array of potentially useful data to be entered, it does not mean that the inclusion of every possible data is useful. The selection of possible predictor variables is often a difficult process and expert knowledge is required. For this purpose, a stepwise linear regression (SLR) method $^{40}$ from S-PLUS $^{41}$, a classification and regression tree software package, was applied and the $C_p$ statistic for each variable was examined. The SLR method selects the best subset of predictor variables to be employed in a regression tree model using a stepwise procedure, which alters the model repeatedly at the previous step by adding or removing predictor variables $^{44}$. The $C_p$ statistic provides a convenient criterion for determining whether a model is improved by adding and dropping the predictor variables $^{34}$. The $C_p$ statistic specifies which predictor variables are significantly relevant to the percent tree cover prediction. The $C_p$ statistic is expressed as $^{30}$:

$$C_p = p + \frac{(n - p)(s^2_p - s^2)}{s^2}$$

where $n$ is the number of observations (number of training data), $p$ is the number of coefficients (number of predictor variables plus one), $s^2_p$ is the mean square error (MSE) of the prediction model, and $s^2$ is the minimum MSE among the possible models, which offer the best estimate of the true error $^{30,34}$. As an example to calculate the $C_p$ statistic for predictor variables of surface reflectances, $n$ is 270 cells (number of training data for developing the model) and $p$ is 85 (surface reflectances of 7 bands of MODIS from December 2002 to November 2003). The best-selected predictor variables were then used for constructing a model in Cubist tools by analyzing the relationships among the data set that contains training data and predictor variables, and created an appropriate regression tree and rule set.

3.1 Reducing cloud contamination and monthly compositing of MODIS 2003 data

Clouds contaminate many original MODIS data; different

Fig. 2 Cloud contamination reduction and regression tree methods applied in this study. Both sets of data (MODIS and QuickBird/ETM) were registered to the same projection in Geographic Latitude/Longitude. After monthly compositing, predictor variables were extracted from MODIS 2003 data. Percent tree cover as predicted variable/training data were extracted from QuikBird data.

Selecting predictor variables using SLR method

Creating tree model using Cubist

Percent tree cover estimation using NLCD mapping tool

Global percent tree cover map

Validation
parts of the same image might be cloud-covered in a series of 16-day data. To remove obvious cloud contamination, two time series data sets were also downloaded: year 2002 and 2004. Thus, for any given pixel, time series MODIS data from 2002 to 2004 were also used in the analyses\(^1\). In the 2003 data, pixels with cloud contamination were then replaced either by 2002 or 2004 data. A linear interpolation was performed for pixels with cloud-cover less than 6 periods. While for pixels with cloud-cover more than 6 periods, those cloud-contaminated pixels (in a specific period) were replaced by averaging MODIS 2002 and 2004, or by either of them if the other was cloudy contaminated for the same band, and then followed by a linear interpolation. This step applied for all 7 bands of MODIS 2003 data (23 periods of 16-day composite data). Although this technique reduces cloud effects, the resulting composite data are usually not entirely cloud-free. For this reason, the post cloud removing of 16-day MODIS 2003 data were then re-composited to 32-day (monthly) data using a MVC formula\(^3\) to reduce the remaining clouds that still appear in these data. A MVC formula applied to the data was created by selecting the maximum value for each pixel in a pair of 16-day composite data. Further compositing of the data set into monthly maximum images produced a data set that appeared to be relatively cloud-free. Although image-compositing of MODIS 1 km to monthly data reduced their temporal resolution, but more complete coverage with relatively cloud-free data was achieved. The steps applied for reducing cloud contamination in this study are shown in Fig. 2.

3.2 Extracting tree and non-tree from QuickBird images

Successful modeling using regression tree techniques relies on the quality of ground truth data. In this study, the percent tree cover as ground truth data was derived from 11 scenes of very-high-resolution QuickBird data. The ground truth data extracted from QuickBird are important because the spatial resolution of those images makes it easy to derive the percent tree and non-tree covers from the image. Those very-high-resolution images also help us to reduce difficulty in defining whether shadowed areas are tree or non-tree. However, the ground truth data extracted from ETM image were used only to represent bare areas where the percent tree cover is zero. A total of 195 cells derived from QuickBird data (each cell corresponding to one MODIS 1 km pixel) and 105 cells from ETM were constructed representing spectral and spatial variability of land cover areas for Africa, ranging from herbaceous to broad-leaved forests. These ground truth data represent values ranging from 0 to 100% tree covers (see Fig. 1 and Table 1).

In this study, unsupervised k-means clustering, supervised maximum likelihood classification and finally on-screen digitizing were employed to extract the percent tree cover from QuickBird images. An unsupervised clustering was the main method for discriminating tree and non-tree due to its capability to automatically distinct classes, which dramatically reduces the work of the analyst\(^3\). For unsupervised clustering, each cluster was interpreted and labeled as a tree or non-tree class. Not all tree-class confusions were modeled successfully. Separation of spectrally similar tree covers and shadowed areas into appropriate tree classes was accomplished using intensive visual interpretation. If the similar tree classes (non-tree classes) and shadowed areas were difficult to be separated, such as in plantations and dense forest areas, then a manual editing and on-screen digitizing were also performed to obtain better class-separation results. For cases in which the result of unsupervised clustering was unsatisfactory, the tree and non-tree extraction was further modified by supervised classification, followed by on-screen digitizing (see Fig. 2). The percent tree cover is then calculated as the percent of tree class pixels for each 1 x 1-km area corresponding to one MODIS pixel. Fig. 3 illustrates an example of how to extract the percent tree cover from QuickBird images using unsupervised clustering method. In all, 300 cells (comprising 195 cells extracted from QuickBird and 105 cells extracted from ETM) were constructed after combining an ETM image. From 300 cells of ground truth data, 270 cells were used as training data and 30 cells were used as validation data. The percent tree cover data as a predicted variable were then integrated to predictor-variables data that were derived from MODIS; both sets of data were used in the regression tree method.

3.3 Regression tree method

Cubist tools, which is one of regression tree algorithm, produce rule-based regression models for prediction of patterns of continuous variables based on the data set. The model can handle a broad range of predictors along with variability of the data set\(^3\), such as variability in surface reflectance recorded by bands and pronological variability in NDVI. The model can therefore be thought of as a piecewise linear model. Each resulting production rule defines conditions under which certain multivariate linear sub-models are applicable in predicting the percent tree cover. Unlike traditional regression tree classifications, these linear models are not mutually exclusive, allowing overlaps among several sub-models. Such piecewise-linear models can account for non-linear multiple relationships between training data and predictor variables\(^3\).\(^3\).

Performance of the Cubist models is measured using the average error, relative error and correlation coefficient (\(r\)). The average error magnitude is the mean absolute error. The relative error magnitude is the ratio of the average error...
magnitude to the error magnitude that would result from always predicting the mean value. The correlation coefficient is a measure of the agreement between actual values of the explanatory variable and those predicted by the model. A correlation coefficient of 1 indicates perfect agreement. An average error $R$ of a tree $T$ is expressed as$^{19}$:

$$ R(T) = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i| $$

(5)

where $\hat{y}_i$ represents the predicted value, $n$ is the number of samples used to build the tree model, and $y$ is the percent tree cover derived from the training data. Whereas the relative error is expressed as$^{19}$:

$$ RE(T) = \frac{R(T)}{R(\mu)} $$

(6)

where $R(\mu)$ is the error magnitude resulting from always predicting the mean value. For useful models, this should be less than 1.

For evaluating the developed model, 26 cells derived from QuickBird and 4 cells derived from ETM were separated from the total ground truth data to validate the model. The rules created from the developed model in Cubist were then extrapolated spatially to the entire 1 km MODIS data to produce a final percent tree cover map of Africa using the NLCD mapping tool (see Fig. 2). The rules were applied according to pixel values within the corresponding input imagery of predictor variables. The NLCD mapping tool was designed for a national land cover project of the United States Geological Survey (USGS)$^{32}$.

4. Results and Discussions

Three regression tree models were built using Cubist tools with combinations of different input variables: all variables (including surface reflectance, NDVI, EVI, NDSI and LST variables), surface reflectance variables, and SLR-selected variables. Surface reflectance variables model contains only the predictor variables of surface reflectance bands of MODIS data. While in SLR-selected variables model, the predictor variables were selected in term of the relevance to the percent tree cover estimation based on its $C_p$ statistics. Higher $C_p$ values compared to their initial $C_p$ model values indicate more relevant variables for prediction (Table 2). Fig. 4 illustrates an example of how to select the predictor variables of surface reflectance that later on will be used as a part of input variables in the SLR-selected variables model. The $C_p$ statistics for surface reflectance variables in the first step before adding and dropping variables are shown in the solid bar graph. The values vary, they are greater or lower than the $C_p$ value of the initial model in this step (see a dotted line). After adding and dropping the variables, the method then produced a new $C_p$ value for a revised model. The predictor variables with $C_p$ values greater than the $C_p$ for the initial model were then included in developing the model using Cubist tools. The same approach was also applied for other predictor variables: NDVI, EVI, NDSI and LST. Table 2 shows the final $C_p$ statistics for all significant variables.

The $C_p$ statistics for the initial model of surface reflectance, NDVI, EVI, NDSI, LST Day and LST Night variables were 12,710, 75,597, 47,644, 83,153, 144,585 and 208,609, respectively. For example, in the surface reflectance variables of
Figure 4: $C_p$ statistic for surface reflectance variables in the first step of SLR method before adding and dropping the variables (solid bar graph). A dotted line shows the $C_p$ value of the initial model in this step. After adding and dropping the variables, the method then produced a new $C_p$ value for a revised model.

Table 2: Selected predictor variables using SLR method and their $C_p$ statistics.

<table>
<thead>
<tr>
<th>No</th>
<th>Predictor Variables</th>
<th>$C_p$ Statistic of Initial Model</th>
<th>Months</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Surface reflectance (12710)</td>
<td>Dec(b1/12924, b2/12774), Jun(b2/12924, b5/12955, b6/12601, b7/13139), Jul(b3/12948, b8/12983, b7/13347), Feb(b5/12859, b6/12860, Mar(b3/12799, b5/12432, b6/12342, b7/13098), Apr(b1/12013, b2/12597), May(b1/12553, b2/12525, b4/12996, b5/13026, b6/13149, b1/12302), Jun(b5/12302, b6/13123), Jul(b2/13681, b4/12994, b7/12972), Aug(b6/13734, b6/12308, b6/12792, b7/13373), Sep(b1/12724), Oct(b1/12724), b6/13681, b6/13584, b7/12972)</td>
<td>Dec(12710), Jun(12710), Jul(12710), Aug(12710)</td>
</tr>
<tr>
<td>2</td>
<td>NDVI (5597)</td>
<td>Jan(b2/5863), Feb(b2/5863), May(b7/4930), Jun(b11776), Aug(b2/5863, Sep(b2/5863), Nov(b2/5863)</td>
<td>Apr(47646), May(47646), Jun(47646), Jul(47646), Aug(47646)</td>
</tr>
<tr>
<td>3</td>
<td>EVI (47646)</td>
<td>Dec(b6/5583), Jan(b7/725), Feb(b4/926), Apr(b4/926), May(b4/926), Jun(b7/725), Aug(b7/1399, Sep(b7/725), Nov(b7/725)</td>
<td>Feb(47646), Mar(47646), Apr(47646), May(47646), Jun(47646), Jul(47646), Aug(47646)</td>
</tr>
<tr>
<td>4</td>
<td>NDSI (83153)</td>
<td>Dec(b6/5583), Feb(b5/4421), Mar(b7/725), Apr(b7/725), May(b7/725), Jun(b7/725), Aug(b7/725)</td>
<td>Sep(b5/4421), Oct(b5/4421), Nov(b5/4421)</td>
</tr>
<tr>
<td>5</td>
<td>LST Day (144585)</td>
<td>Feb(144585), Mar(204873, Apr(17668), May(137834), Jun(167970), Jul(149591), Aug(142873), Sep(178139), Oct(149591), Nov(142873)</td>
<td>Dec(144585), Jan(212688), Mar(204873, Apr(222757), May(212757), Jun(209564), Aug(221374), Sep(209564), Oct(217898)</td>
</tr>
<tr>
<td>6</td>
<td>LST Night (20869)</td>
<td>Dec(129148), Jan(212688), Mar(204873, Apr(222757), May(212757), Jun(209564), Aug(221374), Sep(209564), Oct(217898)</td>
<td></td>
</tr>
</tbody>
</table>

That the tree-model performance had good prediction for estimating the percent tree cover.

The final percent tree cover map, resulting from SLR-selected variables method is shown in Fig. 6. The values range from 0 to 100%. The zero percent tree covers wide areas in northern and southwestern parts of Africa, whereas 100% tree covers distributed near central equatorial Africa. In the

![Diagram of tree model constructed from SLR-selected-variables model](image-url)
Estimating Percent Tree Cover Using Regression Tree Method with Very-High-Resolution QuickBird Images as Training Data

Fig. 5  Tree model constructed from SLR-selected variables. Rectangles show the terminal node with its percent tree cover values resulted from training data. Variable used for creating the regression tree model are shown in ellipses. This figure shows only 5 top splits from the total 12 splits.

Fig. 6  Final percent tree cover map for Africa using regression tree method. The dark green areas represent those with high percentage of tree cover and white areas represent those with zero percent tree cover.

northwestern part of Africa near coast areas, the percentage ranges from 0 to 30%. A high percentage of tree cover distributed also in the southeastern Africa and the eastern to southeastern parts of Madagascar. In the southern part, the percentage ranges from 0 to 40%. This map also shows percent tree cover for areas that are frequently covered by clouds in raw MODIS data, e.g., in the western part of Africa and in central Africa (near the equatorial). Using regression
tree method, such areas were resolved based on surface reflectance values from the other cloud-free months. This advantage is important for frequent global forest/tree cover change studies.

Validation of the regression tree result was made through validation cells derived from QuickBird and ETM images. The validation cells were selected from ground truth data based on a stratified random sampling method and also contain percent tree cover values of 0 to 100%. In this sampling method, 5–10 percent of the total cells in each stratum were selected. Table 3 shows the prediction error of different variable combinations. Comparison of the prediction errors in this table revealed that the prediction model derived from SLR-selected variables produced the lowest prediction error (5.17%) and contributes more significant in depicting the percent tree cover than the other two models. This model also indicates that the prediction model with all variables, with prediction error of 7.17%, did not produce better results for differentiating the percent tree cover than the SLR-selected model. Some reflectance bands did not markedly improve the prediction results of percent tree cover. Table 3 also shows that SLR-selected variables produced a higher correlation coefficient (0.962) than the all-variables model (0.937), whereas the surface reflectance model produced a slightly lower correlation coefficient (0.960). The correlation coefficient of the SLR-selected-variables model is also illustrated in the scatter plot, which depicts this model result as opposed to the ground truth (Fig. 7 (a)). This scatter plot shows good agreement to the predicted percent tree cover with $r = 0.962$. This value shows that the SLR-selected variables contribute considerably to improve the regression tree method results.

Comparison with the percent tree cover produced by UM[2] was also examined to check accuracy resulted from both results (Table 3). The prediction error of the SLR-selected variables' result is 5.17% while the result of UM is 10.43%. Fig. 7 (b) shows scatter plots for percent tree result from this study after applying nearest neighbor resampling to 500 m resolution and the UM result. For percent tree cover lower than 30%, the result of this study agrees well with the UM result, for greater percent tree cover the UM result shows worse agreement. Results of this study show that the accuracy from SLR-selected method would be better than results of UM. The training data extracted from QuickBird images differentiated tree and non-tree areas better than training data from MSS/TM data, as tested in this study.

Table 3 Prediction errors and correlation coefficients for all, surface reflectance, and SLR-selected variables. This table also compares the result of this study and the result of UM[2].

<table>
<thead>
<tr>
<th>No.</th>
<th>Predictor Variables</th>
<th>Prediction Error (%)</th>
<th>Correlation Coefficients ($r$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>All variables</td>
<td>7.17</td>
<td>0.937</td>
</tr>
<tr>
<td>2</td>
<td>Surface reflectance variables</td>
<td>5.59</td>
<td>0.960</td>
</tr>
<tr>
<td>3</td>
<td>SLR-selected variables</td>
<td>5.17</td>
<td>0.962</td>
</tr>
<tr>
<td>4</td>
<td>UM result</td>
<td>10.45</td>
<td>0.817</td>
</tr>
</tbody>
</table>

Fig. 7 Scatter plots depict a comparison between the predicted result produced by a regression tree in this study and validation data derived from QuickBird (a). Some points coincide with MODIS 2003 data that had been contaminated by clouds during 1 to 2 months. Comparison between the result of this study and the result of the University of Maryland (UM) are shown in (b). For percent tree cover lower than 40%, the result of this study agrees with the UM[2] result, while for greater percent tree cover, the UM result shows worse agreement.
5. Conclusions

It is demonstrated that QuickBird data give an improvement for depicting the continuous percent tree cover of training data. Accuracy assessment showed that the SLR-selected variables produced the best result, with prediction error of 5.17%. As illustrated in the scatter plot, this study also demonstrates that SLR-selected variables yielded a higher correlation coefficient (0.962) than the all-variables model and the surface reflectance model. The tree model explains that NDVI and surface reflectance bands contribute more significant than LST Day, LST Night, NDSI and EVI to differentiate percent tree cover. The zero-percent tree cover area includes wide areas in northern and southwestern parts of Africa, whereas 100% tree cover was distributed near central equatorial Africa. High percentages of tree cover were distributed also in the southeastern Africa and eastern to southeastern Madagascar. The regression tree is an advance as well; it can handle the non-linear relationships in the training data. Regression tree method is also more robust to overcome cloud problems that exist in most satellite remote sensing data.

6. Acknowledgments

The MODIS data used in this study were acquired as part of NASA’s Earth Science Enterprise. The data were processed using the MODIS Adaptive Processing System (MODAPS) and Goddard Distributed Active Archive Center (DAAC), and are archived and distributed by the Goddard DAAC. The NLCD mapping tool software used courtesy of USGS. The authors would like to thank two anonymous reviewers whose comments improved the quality of the paper. This study was carried out as a part of activities by Working Group 4 of the Global Mapping project.

References

16) UNEP (United Nations Environment Programme), Environment action plan for the implementation of the environment initiative of the new partnership for Africa development, Working draft 1, Nairobi, Kenya, 26, 2002.


Estimating Percent Tree Cover Using Regression Tree Method with Very-High-Resolution QuickBird Images as Training Data

Rokhmatuloh received the Bachelor degree from the Department of Geography, University of Indonesia in 1996. He received a M.Eng. degree in remote sensing from Chiba University, Japan in March 2004. In April 2004, he entered a Ph.D. course at that University. He has attended several local and international conferences during his period of study in Japan. His research interests are data fusion and tree cover mapping for regional and global areas. E-mail: rahmat@graduate.chiba-u.jp

Daisuke Nitto received his Master’s degree in Engineering from Chiba University, Japan in 2005. He received his Bachelor degree from that University. His graduation thesis topic is percent tree cover estimation.

Hussam Al Bilbisi is a researcher at Center for Environmental Remote Sensing, Chiba University. He received his Ph.D. degree in Remote Sensing from Chiba University, Japan in 2004. His research interests include image processing, land remote sensing, change detection, and land degradation using microwave and optical remote sensing data.

Kota Arihara received his Bachelor degree from Chiba University, Japan in March 2006. He wrote his graduation thesis on the topic of MODIS data processing for land cover analysis. He is also holding interests in computer programming.

Ryutaro Tateishi received his Doctor of Engineering degree from the University of Tokyo in 1982. He is a professor at the Center for Environmental Remote Sensing (CEReS), Chiba University. His research activities are in the field of land remote sensing, and are currently focused on global land cover mapping and monitoring. He was a Secretary of ISPRS Commission IV during 1988–1992, and the Chairman of ISPRS WG IV/6 and IV/8, respectively, during 1992–2000 and 2000–2004. He was also the Chairman of Land Cover WG of AARS during 1993–1999. He is currently the Chairman of International Steering Committee for Global Mapping (ISCGM) WG 4 on raster data development since 2002.