Integrated Use of Existing Global Land Cover Datasets for Producing a New Global Land Cover Dataset with a Higher Accuracy: A Case Study in Eurasia

Naijia Zhang*, Ryutaro Tateishi
Center of Environmental Remote Sensing (CEReS), Chiba University, Chiba, Japan
Email: *Z_N_J@graduate.chiba-u.jp

Received October 28, 2013; revised November 28, 2013; accepted December 5, 2013

ABSTRACT

It has been commonly acknowledged that the current global mapping projects have encountered the accuracy challenge. By conducting a comparison among the four existing global land cover datasets (MODIS LC, GLC2000, GLCNMO and GLOBCOVER), it has been identified that certain areas’ accuracy has dragged down the overall accuracy of these global land cover datasets. In this paper, those areas have been defined as the “unreliable area”. This study has recollected the training data from the “unreliable area” within the above four mentioned datasets and reclassified the “unreliable area” by using two supervised classifications. The final result has shown that compared with any existing datasets, a relatively higher accuracy has been able to achieve.

Keywords: Global Land Cover; GLCNMO; Training Data; Accuracy

1. Introduction

Global mapping plays an important role in the areas such as monitoring the major environmental phenomena, environmental protection as well as sustainable growth. An accurate global map could also contribute to the establishment of a global spatial data infrastructure, for future research and many other scientific purposes.

Until present, many global land cover projects have been carried out. Examples are that the IGBP DISCover dataset was based on the Advanced Very High Resolution Radiometer (AVHRR) from 1992 to 1993 [1], and the land cover product of the University of Maryland (UMD) was based on the same data from AVHRR, distinguished 14 classes [2]. In 2002, Boston University produced the MODIS land cover data using MODIS 1-km satellite data on board the Terra satellite [3]. The Global Land Cover 2000 (GLC2000) was based on SPOT-VEGETATION data from November 1999 to December 2000 [4,5]. Global Land Cover by National Mapping Organizations (GLCNMO) was based on 2003 data from MODIS, which was produced by the Center for Environmental Remote Sensing (CEReS, Chiba University) [6]. In 2009, cooperating with an international network of partners (including EEA, FAO, GOFC-GOLD, IGB, JRC and UNEP), the European Space Agency (ESA) produced GLOBCOVER. Unlike other datasets, GLOBCOVER presents a higher resolution (300 m) than any previous global satellite derived maps [7].

Besides many studies on a single dataset, various researches have also tried to compare the existing different global land cover datasets. In 2006, a spatial comparison of four satellite derived 1 km global land cover datasets (IGBP, UMD, MODIS LC, GLC2000) was conducted by generalizing a global land cover legend [8]. Another comparison between the exiting 1 km datasets was conducted in 2008 [9]. Purpose of those comparisons is trying to develop the integrated use of different datasets. For example, areas having the high agreement from the various existing global datasets were to be served as the reference data for training area selections by Chandra Giri et al.’s study in 2005 [10].

However, the integrated uses so far have mostly focused on the areas with high accuracy. There are large areas with low accuracy, which seem to have been ignored. If the accuracy of these areas could be improved to a higher level, theoretically a better global land cover datasets can be expected and the potential usage can be discovered within those accuracy-improved areas. There-
fore, a question of “How to improve the accuracy level of certain areas” has been raised, which is also the key objective of this paper.

2. Methodology

As Figure 1 shown, this study has utilized the four global land cover datasets: 1) MODIS LC (v004), 2) GLC2000 (v1.1), 3) GLCNMO (2003) and 4) GLOBCOVER (2009). The detail information of classes of each datasets is provided in Appendix.

This study used these datasets to separate the high accuracy area and the low accuracy area. Next, for the reclassification purpose, the low accuracy area has been checked cautiously to collect the training data. Two classification methods (Maximum likelihood method and decision tree method) have been adopted to produce the accuracy result as well as to compare. Finally, the accuracy comparison has been done between the results and the existing datasets.

2.1. Preprocessing

As mentioned above, there is a resolution difference between MODIS LC (v004), GLC2000 (v1.1), GLCNMO (2003) and GLOBCOVER (2009). Therefore, to be able to compare, the first step was to resample them all to the same resolution, which was a 300 m resolution same as GLOBCOVER (2009).

Next step was to reconcile the different legends (Table 1), again due to the differences among those four datasets. Most classes (i.e. some part of the forest, urban, bare land and water bodies etc.) were translated well. However, the “mixed classes” were difficult to correspond with each other. In this study, the correspondences were mainly based on the GLCNMO’s classes [11-17].

Table 1 shows the pixel-by-pixel comparison of four maps.

2.2. Area Separation Based on the Accuracy Assessment

The information provided by four global land cover datasets could lead to four levels of synthesized agreements, which are listed as below:

Zone 1: No agreement in all datasets.

Zone 2: The first two datasets are in agreement and the other two are also in agreement.

Another situation is only two of the four datasets are in agreement while the other two are not.

Zone 3: Agreement among three datasets.

Zone 4: Agreement among all the four datasets.

According to the above information, the regions of Zone 3 and Zone 4 are defined as the “reliable area” (Figure 2) in this study. Consequently, the regions of Zone 1 and Zone 2 (as the blank part of Figure 2) are defined as the “unreliable area”.

Regarding those so called the “reliable area”, are they truly reliable (with highly accuracy)? To confirm those zones have been defined correctly, an accuracy assessment was conducted. The classes with the majority agreements were adopted directly in zone 3 and zone 4. On the other hand, it was difficult to decide the certain classes based agreements in zone 1 and zone 2. Therefore, the blank parts were filled with the GLCNMO’s classes.

A total number of about 1800 validation points were

![Figure 1. Methodology.](image-url)

### Table 1. The seventeen aggregated classes for the four land cover products.

<table>
<thead>
<tr>
<th>New code</th>
<th>1) GLCNMO</th>
<th>2) MODIS LC</th>
<th>3) GLC2000</th>
<th>4) GLOBCOVER</th>
</tr>
</thead>
<tbody>
<tr>
<td>1) Broadleaf Evergreen Forest</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>40</td>
</tr>
<tr>
<td>2) Broadleaf Deciduous Forest</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>50</td>
</tr>
<tr>
<td>3) Needleleaf Evergreen Forest</td>
<td>3</td>
<td>1</td>
<td>4</td>
<td>70</td>
</tr>
<tr>
<td>4) Needleleaf Deciduous Forest</td>
<td>4</td>
<td>3</td>
<td>5</td>
<td>N/A</td>
</tr>
<tr>
<td>5) Mixed Forest</td>
<td>5</td>
<td>5</td>
<td>6</td>
<td>100</td>
</tr>
<tr>
<td>6) Tree Open</td>
<td>6</td>
<td>8</td>
<td>3</td>
<td>60, 90</td>
</tr>
<tr>
<td>7) Shrub</td>
<td>7</td>
<td>6, 7</td>
<td>11, 12</td>
<td>110, 120, 130</td>
</tr>
<tr>
<td>8) Herbaceous</td>
<td>8, 9</td>
<td>9, 10</td>
<td>13</td>
<td>140</td>
</tr>
<tr>
<td>9) Sparse Vegetation</td>
<td>10</td>
<td>N/A</td>
<td>14</td>
<td>150</td>
</tr>
<tr>
<td>10) Cropland</td>
<td>11, 12</td>
<td>12</td>
<td>16</td>
<td>11, 14</td>
</tr>
<tr>
<td>11) Cropland /other vegetation mosaic</td>
<td>13</td>
<td>14</td>
<td>17, 18</td>
<td>20, 30</td>
</tr>
<tr>
<td>12) Mangrove</td>
<td>14</td>
<td>N/A</td>
<td>8</td>
<td>170</td>
</tr>
<tr>
<td>13) Wetland</td>
<td>15</td>
<td>11</td>
<td>7, 15</td>
<td>160, 180</td>
</tr>
<tr>
<td>14) Bare area</td>
<td>16, 17</td>
<td>16</td>
<td>19</td>
<td>200</td>
</tr>
<tr>
<td>15) Urban</td>
<td>18</td>
<td>13</td>
<td>22</td>
<td>190</td>
</tr>
<tr>
<td>16) Snow/Ice</td>
<td>19</td>
<td>15</td>
<td>21</td>
<td>220</td>
</tr>
<tr>
<td>17) Water Bodies</td>
<td>20</td>
<td>0</td>
<td>20</td>
<td>210</td>
</tr>
</tbody>
</table>
Figure 2. “Reliable area” extracted from four land cover datasets (MODIS LC, GLC2000, GLCNMO, GLOBCOVER).

<table>
<thead>
<tr>
<th></th>
<th>Land Cover Type</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Broadleaf evergreen forest</td>
<td>10</td>
</tr>
<tr>
<td>2</td>
<td>Broadleaf deciduous forest</td>
<td>11</td>
</tr>
<tr>
<td>3</td>
<td>Needleleaf evergreen forest</td>
<td>12</td>
</tr>
<tr>
<td>4</td>
<td>Needleleaf deciduous forest</td>
<td>13</td>
</tr>
<tr>
<td>5</td>
<td>Mixed forest</td>
<td>14</td>
</tr>
<tr>
<td>6</td>
<td>Tree open</td>
<td>15</td>
</tr>
<tr>
<td>7</td>
<td>Shrub</td>
<td>16</td>
</tr>
<tr>
<td>8</td>
<td>Herbaceous</td>
<td>17</td>
</tr>
</tbody>
</table>

1) Satellite image of Google Earth.
2) Ground photographs near the locations in Google Earth.
3) Ground photographs of the Degree Confluence Project (http://confluence.org/).

Out of which, about 800 validation points were successfully identified as shown in Figure 3.

The final validation result is shown in the Table 2 below.

The final result has shown an average accuracy of approximately 76%, which is generally same as the overall accuracy of the existing global land cover datasets. Similar tests were conducted as to compare, i.e. filled the blank parts with other global land cover datasets (MODIS LC, GLC2000, GLOBCOVER) and the similar results were achieved.

As the validation result, simply by overlaying the existing global land cover dataset, the overall accuracy cannot be improved. At the same time, it has also revealed the “unreliable area” has dragged down the overall accuracy of these global land cover datasets. Many factors could lead to the appearance of the “unreliable area”, and examples are the complexity of the geographic systems, the different resolutions and resources of satellite data, and the different definitions of classes etc.

Another critical factor that leads to the appearance of the “unreliable area” is the different classification methods that being adopted in different land cover datasets. Among all methods, supervised classification is mostly commonly adopted. During the supervised classification processing, the quality of training data plays an essential role. Therefore one assumption has been proposed, which is the lack of quality training data that caused the “unreliable areas” (zone 1 and zone 2). To verify such assumption, the training data of GLCNMO was double-checked. The result has shown that most training data in GLCNMO was generated from the “reliable area”, thus the assumption has been verified.

2.3. Recollection of Training Data and Reclassification

To be able to reclassify, this paper has used MODIS 2008 16-day composite imagery (http://glcf.umd.edu/data/modis/). Center for Environment...
mental Remote Sensing (CEReS, Chiba University) has completed the pre-processing. MODIS 5 bands (band 1, 2, 4, 6, 7) and derived Normalized Difference Vegetation Index (NDVI) were used for the classification. 

Figure 4 shows the case study area in this paper. Eleven land cover classes indicated in Table 3 were classified by the supervised classification. On the contrast, the other six land cover classes were difficult to be determined by the supervised method according to the GLCNMO experiences. The training data were collected from the “unreliable area”, which was checked cautiously to ensure the quality. First of all, for proper program processing, the number of pixels should not be less than 72 in each sub-class. If there was no sufficient training data at the “unreliable area”, the training data from the “reliable area” that has the same characteristics was adopted. All the training data (every pixel) in this study were added, deleted or modified, according to the MODIS 2008’s NDVI seasonal (23 periods) patterns.

As the end result, eleven land cover classes have been divided into 81 sub-classes (Table 4).

### 3. Reclassification Result

Maximum likelihood method (MLC) by ENVI software was adopted, similar to the previous GLCNMO project. Decision tree method (DCT) by See5 software and CART software was used as well.

The classification result is shown in Figure 5 and Figure 6.

### 4. Accuracy Comparison

In order to validate the results, another total number of 800 validation points was further taken randomly for the 11 classes. As the same identify method mentioned above, about 400 validation points were successfully identified.

Figure 7 shows the accuracy comparison between the two results (as per Figures 5 and 6) and the four existing datasets. The average accuracy of the existing global land cover datasets is approximately 56%. Relatively, the average accuracy of result 1 is 70.13% and result 2 is 65.93%.

### 5. Conclusion

Using the existing land cover datasets or the existing local data products does make the training data preparation more efficient. While such methods tend to extract the training data mostly from the “reliable area”, this study has proved that the training data collected from “unreliable area” are very important as well. This paper shows that the accuracy of “unreliable area” can be

---

**Table 2. The accuracy of each zone.**

<table>
<thead>
<tr>
<th>Area</th>
<th>Zone 4</th>
<th>Accuracy (%)</th>
<th>Zone 3</th>
<th>Accuracy (%)</th>
<th>Zone 2</th>
<th>Accuracy (%)</th>
<th>Zone 1</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reliable area</td>
<td>98.16</td>
<td>93.59</td>
<td>89.02</td>
<td>75.89</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unreliable area</td>
<td>63.13</td>
<td>58.18</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 3. Land cover classes that were classified by supervised classification method.**

<table>
<thead>
<tr>
<th>New code</th>
<th>Collection of training data</th>
</tr>
</thead>
<tbody>
<tr>
<td>1) Broadleaf Evergreen Forest</td>
<td>○</td>
</tr>
<tr>
<td>2) Broadleaf Deciduous Forest</td>
<td>○</td>
</tr>
<tr>
<td>3) Needleleaf Evergreen Forest</td>
<td>○</td>
</tr>
<tr>
<td>4) Needleleaf Deciduous Forest</td>
<td>○</td>
</tr>
<tr>
<td>5) Mixed Forest</td>
<td>○</td>
</tr>
<tr>
<td>6) Tree Open</td>
<td>-</td>
</tr>
<tr>
<td>7) Shrub</td>
<td>○</td>
</tr>
<tr>
<td>8) Herbaceous</td>
<td>○</td>
</tr>
<tr>
<td>9) Sparse Vegetation</td>
<td>○</td>
</tr>
<tr>
<td>10) Cropland</td>
<td>○</td>
</tr>
<tr>
<td>11) Cropland/other vegetation mosaic</td>
<td>○</td>
</tr>
<tr>
<td>12) Mangrove</td>
<td>-</td>
</tr>
<tr>
<td>13) Wetland</td>
<td>-</td>
</tr>
<tr>
<td>14) Bare area</td>
<td>○</td>
</tr>
<tr>
<td>15) Urban</td>
<td>-</td>
</tr>
<tr>
<td>16) Snow/ice</td>
<td>-</td>
</tr>
<tr>
<td>17) Water Bodies</td>
<td>-</td>
</tr>
</tbody>
</table>
Figure 5. Result 1 (by MLC).

Figure 6. Result 2 (by DCT).
Table 4. Number of land cover sub-classes.

<table>
<thead>
<tr>
<th>New code</th>
<th>Number of sub-classes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1) Broadleaf Evergreen Forest</td>
<td>6</td>
</tr>
<tr>
<td>2) Broadleaf Deciduous Forest</td>
<td>8</td>
</tr>
<tr>
<td>3) Needleleaf Evergreen Forest</td>
<td>6</td>
</tr>
<tr>
<td>4) Needleleaf Deciduous Forest</td>
<td>8</td>
</tr>
<tr>
<td>5) Mixed Forest</td>
<td>5</td>
</tr>
<tr>
<td>6) Tree Open</td>
<td>-</td>
</tr>
<tr>
<td>7) Shrub</td>
<td>8</td>
</tr>
<tr>
<td>8) Herbaceous</td>
<td>12</td>
</tr>
<tr>
<td>9) Sparse Vegetation</td>
<td>8</td>
</tr>
<tr>
<td>10) Cropland</td>
<td>24</td>
</tr>
<tr>
<td>11) Cropland/other vegetation mosaic</td>
<td>8</td>
</tr>
<tr>
<td>12) Mangrove</td>
<td>-</td>
</tr>
<tr>
<td>13) Wetland</td>
<td>-</td>
</tr>
<tr>
<td>14) Bare area</td>
<td>8</td>
</tr>
<tr>
<td>15) Urban</td>
<td></td>
</tr>
<tr>
<td>16) Snow/Ice</td>
<td></td>
</tr>
<tr>
<td>17) Water Bodies</td>
<td></td>
</tr>
</tbody>
</table>

Figure 7. The accuracy comparison between the New Map 1, 2, and the existing maps (Eurasia).

improved to a higher level, comparing to the previous studies. Such results were achieved by recollecting the quality training data and reclassification. Although only 11 classes from the “unreliable area” were classified in this study, it has focused on the most unreliable area of the map and a minimum of 65.93% was achieved. In other words, if such methodology could be introduced to the other continents, a global land cover dataset that has a higher accuracy can be expected.

REFERENCES


[12] M. Herold, R. Hubald and A. Di Gregororio, “Translating and Evaluating Land Cover Legends Using the UN Land


Appendix

**MODIS**

0) Water Bodies
1) Evergreen Needleleaf Forests
2) Evergreen Broadleaf Forests
3) Deciduous Needleleaf Forests
4) Deciduous Broadleaf Forests
5) Mixed Forests
6) Closed Shrublands
7) Open Shrublands
8) Woody Savannas
9) Savanna
10) Grasslands
11) Permanent Wetlands
12) Croplands
13) Urban and Built-up
14) Cropland/Natural Vegetation Mosaics
15) Snow and Ice
16) Barren

**GLC2000**

1) Tree Cover, broadleaved, evergreen
2) Tree Cover, broadleaved, deciduous, closed
3) Tree Cover, broadleaved, deciduous, open
4) Tree Cover, needle-leaved, evergreen
5) Tree Cover, needle-leaved, deciduous
6) Tree Cover, mixed leaf type
7) Tree cover, regularly flooded, fresh water
8) Tree Cover, regularly flooded, saline water
9) Mosaic: Tree Cover/Other natural vegetation
10) Tree Cover, burnt
11) Shrub Cover, closed-open, evergreen
12) Shrub Cover, closed-open, deciduous
13) Herbaceous Cover, closed-open
14) Sparse Herbaceous or sparse Shrub Cover
15) Regularly flooded Shrub and/or Herbaceous Cover
16) Cultivated and managed areas
17) Mosaic: Cropland/Tree Cover/Other natural vegetation
18) Mosaic: Cropland/Shrub or Grass Covered
19) Bare Areas
20) Water Bodies
21) Snow and Ice
22) Artificial surfaces and associated areas

**GLCNMO**

1) Broadleaf Evergreen Forest
2) Broadleaf Deciduous Forest
3) Needleleaf Evergreen Forest
4) Needleleaf Deciduous Forest
5) Mixed Forest
6) Tree Open
7) Shrub
8) Herbaceous, single layer
9) Herbaceous with Sparse and Tree/Shrub
10) Sparse Vegetation
11) Cropland
12) Paddy field
13) Cropland/Natural Vegetation Mosaic
14) Mangrove
15) Wetland
16) Bare area, consolidated (gravel, rock)
17) Bare area, unconsolidated (sand)
18) Urban
19) Snow/Ice
20) Water Bodies

**GLOBCOVER**

11. Post-flooding or irrigated crop lands
14. Closed to open (>15%) grass land
20. Mosaic cropland (50% - 70%)/vegetation (grass land/shrub land/forest) (20% - 50%)
30. Mosaic vegetation (grass land/shrub land/forest) (50% - 70%) /cropland (20% - 50%)
40. Closed to open (>15%) broad leaved evergreen and/or semi-deciduous forest (>5 m)
50. Closed (<40%) broad leaved deciduous forest (>5 m)
60. Open (15% - 40%) broad leaved deciduous forest/wood land (>5 m)
70. Closed (>40%) needle-leaved evergreen forest (>5 m)
80. Open (15% - 40%) needle-leaved deciduous or evergreen forest (>5 m)
100. Closed to open (>15%) mixed broadleaved and need leaved forest
110. Mosaic forest or shrub land (50% - 70%) and grass land (20% - 50%)
120. Mosaic grassland (50% - 70%) and forest or shrub land (20% - 50%)
130. Closed to open (>15%) shrub land (<5 m)
140. Closed to open (>15%) grass land
150. Sparse (<15%) vegetation
160. Closed (>40%) broad leaved forest regularly flooded, fresh water
170. Closed (>40%) broadleaved semi-deciduous and/or evergreen forest regularly flooded, saline water
180. Closed to open (>15%) grassland or woody vegetation on regularly flooded or waterlogged soil—Fresh, brackish or saline water
190. Artificial surfaces and associated areas (Urban areas >50%)
200. Bare areas
210. Water bodies
220. Permanent Snow and Ice