Decision tree classification of land use land cover for Delhi, India using IRS-P6 AWiFS data

Milap Punia, P.K. Joshi, M.C. Porwal

Centre for the Study of Regional Development, Jawaharlal Nehru University, New Delhi 110067, India
Department of Natural Resources, TERI University, New Delhi 110070, India
Forestry and Ecology Division, Indian Institute of Remote Sensing, Dehradun 248001, India

Abstract

In this study we explored the potential of multi-temporal IRS P6 (Resourcesat) Advanced Wide Field Sensor (AWiFS) data for mapping of LULC for Delhi, India. The study presents the result of a decision tree classification of seasonal composite data (three seasons). The study has identified 13 classes with description of cropping pattern namely, double crops, kharif, rabi and zaid from 56 m spatial resolution AWiFS data. Delhi has a diverse range of land use predominantly mosaic of built-up. More than half of the area is urban settlement. Results indicate that the temporal data set with a good definition of training sites can result in good overall accuracy (=91.81) as well as individual classification accuracies (producers accuracy >76.92 and users accuracy >60). It is evident that AWiFS data can be used to provide timely and detailed LULC maps with limited ancillary data. The AWiFS derived maps could be very useful as input to biogeochemical models that require timely estimation of LULC patterns.

Keywords: LULC, Resourcesat, AWiFS, Decision tree approach, Multitemporal

1. Introduction

Remote sensing technology is one proven strategy to better document, characterise and quantify land use land cover (LULC) (Wentz, Nelson, Rahman, Stefanov, & Roy, 2008). This information is vital input for various developmental, environmental and resource planning applications, and regional as well as global scale process models. These kinds of databases are also important for national accounting of natural resources and planning at regular intervals. Furthermore, the spatial information addressing the kharif (June–November), rabi (October–March) and zaid (March–July) crops, greening of wastelands, seasonal dynamics of wetlands, surface water bodies, forest vegetation, urbanisation and other high temporal land use practices using satellite remote sensing data can provide a reliable database in timely manner.

The challenge in classifying LULC in urban areas with multi-spectral data is that urban landscape is heterogeneous (Stefanov, Ramsey, & Christensen, 2001). Mixed pixel is a common confounding factor in classification using moderate resolution datasets (Small, 2003; Woodcock & Strahler, 1987). To resolve these issues, investigators have utilise various approaches like, artificial neural networks (Pu, Gong, Michishta, & Sasgawa, 2008), fuzzy classifier (Feitosa, Costa, Mota, Pakzad, & Costa, 2009; Thapa & Murayama, 2009), image segmentation (Gamanya, Maeyer, & Dapper, 2007), expert classification (Stefanov et al., 2001; Wentz et al., 2008), super vector machines (Carräa, Goncalves, & Caetano, 2008) and many others. A moderate resolution data with high temporal and radiometric dimensionality over a large area, certainly calls for huge computing resources as well as robust classification procedures to resolve the challenges of classification in urban area. Further, the classification procedures should have the ability to handle the temporal spectral variability to capture the information on various LULC classes. The extraction of ‘land use’ information from satellite data is often difficult since it is closely associated with the human intervention for which the data need to be obtained from other sources. Hence, the classifier should not make any a priori assumption about the data distribution (Joshi, Roy, Singh, Agrawal, & Yadav, 2006; Kandrika & Roy, 2008).

Keeping all the requirements and constraints in view, a non-parametric approach that can train quickly with a capability to handle huge data sets from numeric and non-numeric sources is a good choice. In this context, decision tree classification techniques have been successfully used for digital image classification (Kandrika & Roy, 2008; Sesnie, Gessler, Finegan, & Thessler, 2008; Tooke, Coops, Goodwin, & Voogt, 2009). This is suitable for remote sensing classification problems because of its flexibility, intuitive simplicity and computational efficiency, which led to increased acceptance (Pal & Mather, 2003; Quiland, 1993). A variety of works have demonstrated that decision tree provides an accurate and efficient methodology for land cover classification (DeFries, Hanson, Townshend, & Sohlmberg, 1998; Friedl, Brodley, & Strahler,
1999; Hansen, DeFries, Townshend, & Sohlberg, 2000). The advantages of using decision tree classification approach are (i) ability to handle noisy and missing data; (ii) require no assumptions regarding the distribution of input data; and (iii) provide an intuitive classification structure.

In this study, we present the results of adapting a decision tree approach for classifying LULC to the metropolitan area of Delhi, India. The multi-fold objectives of the study aim to assess (i) how well decision tree approach works, (ii) potential of multi-temporal IRS P6 (Resourcesat) Advanced Wide Field Sensor (AWiFS) for LULC classification and (iii) to use classified image to document the net sown area for Delhi, India. This type of documentation provides support for human impact investigation related to ecological change, socio-economic interferences, infrastructure development and human health.

1.1. IRS P6 AWiFS data

IRS P6 is the tenth satellite of ISRO in IRS series. It is intended to continue the remote sensing data services provided by IRS-1C and IRS-1D, both of which have far outlived their designed mission lives, and also enhance the data quality. The AWiFS instrument is a space borne optical sensor that is designed for observation of vegetation and land surfaces. It is an improved version compared to the WiFS sensor flown in IRS-1C/1D. It operates in three spectral bands in VNIR (B2 (green, 0.52–0.59 μm), B3 (red, 0.62–0.68 μm), B4 (NIR, 0.77–0.86 μm)) and one band in SWIR (1.55–1.70 μm). With a swath of 740 km, AWiFS provides temporal resolution of 5 days at 56 m spatial and 10 bit radiometric resolution. Compared with AVHRR, SPOT and MODIS, and WiFS that are used for vegetation cover monitoring, AWiFS provides better spatial resolution with acceptable temporal resolution. In this study, a datasets of three seasons in 2004/2005 were used (Table 1). An IRS-P6 AWiFS False Color Composite (FCC) of Delhi is shown in Fig. 1.

2. Study area

The study area covers the administrative boundary of Delhi. It lies between the latitudinal parallels of 28°40’ N and 28°67’ and the longitudinal parallels of 77°14’ E and 77°22’ E (Fig. 2, Delhi region) and occupies northern region of India (216 m above sea level). With an area of 1483 sq. km, it corresponds to a typical patch of the tropical region, completely engrossed with residential, commercial and urban centers. Its South–north length is approximately 54 km and East–west distance is 51 km. The climate is classified as continental because of its distance from the sea with temperature range varying from 45°C in summers to 4°C in winters, rains (annual average rainfall 750–1500 mm) are spread throughout the year. Fig. 2 shows the location of the study area.

3. Materials and methods

3.1. Remote sensing data

The data used for the study is multi-temporal AWiFS of six different dates spread over one crop cycle. Keeping in view the

Table 1
Satellite data for different seasons.

<table>
<thead>
<tr>
<th>S. No.</th>
<th>Season</th>
<th>Dates of acquisition</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Kharif (Monsoon season)</td>
<td>28 August 2005</td>
</tr>
<tr>
<td></td>
<td></td>
<td>15 October 2004</td>
</tr>
<tr>
<td>2.</td>
<td>Rabi (Winter Season)</td>
<td>15 January 2005</td>
</tr>
<tr>
<td></td>
<td></td>
<td>13 March 2005</td>
</tr>
<tr>
<td>3.</td>
<td>Zaid (Summer season)</td>
<td>01 April 2005</td>
</tr>
<tr>
<td></td>
<td></td>
<td>14 May 2005</td>
</tr>
</tbody>
</table>

Fig. 1. False color composite of IRS P6 AWiFS data showing Delhi, January 2005.

heterogeneity of crop calendar over the entire region, even within a season, multidate satellite data is used. False Color Composite (FCC) of Delhi is shown in Fig. 1. Table 1 lists satellite data pertaining to Delhi state from August 2004 to May 2005 to account for cropped area inclusive of other LULC classes.

3.2. Data pre-processing

The image processing software ERDAS Imagine was used. The satellite data were geometrically registered using second order polynomial transformation. The uniformly distributed GCPs were used in such a way that the RMS error is less than 0.33 pixel. The resample (using nearest neighborhood algorithm) data set was overlaid on the reference data and checked for image to image matching. All the scenes were then co-registered for further analysis. The study area was extracted using digital boundary data provided by Survey of India (SOI). Erdas Imagine 8.6 was used for the data processing and further LULC mapping. For radiometric correction, ‘Top-of-atmosphere’ (TOA) reflectance was calculated based on a physic mode and AWiFS sensor calibration factors. The gain and offset calibration were applied using the gain and offset data provided for each image band. These values for each image band were obtained from AWiFS sensor report. The sun zenith angle for each pixel and the distance from the scene center to the sun was calculated first, and then the reflectance correction was calculated for each band. Conversion from calibrated digital numbers \( Q_{\text{cal}} \) in L1 products back to at-sensor radiance \( L \) was carried out using following equation.

\[
L(j) = \frac{(L_{\text{Max}(j)} - L_{\text{Min}(j)})}{Q_{\text{cal Max}}} Q_{\text{cal}} + L_{(j)}
\]

where, \( L_{(j)} \) = Spectral radiance at the sensor's aperture in W/(m² sr µm), \( L_{\text{Max}(j)} \) = Spectral radiance that is scaled to \( Q_{\text{cal Max}} \) in W/(m² sr µm), \( L_{\text{Min}(j)} \) = Spectral radiance that is scaled to \( Q_{\text{cal Min}} \) in W/(m² sr µm), \( Q_{\text{cal}} \) = Quantized calibrated pixel value in digital number, \( Q_{\text{cal Min}} \) = the maximum quantized calibrated pixel value (DN = 0) corresponding to \( L_{\text{Min}(j)} \), \( Q_{\text{cal Max}} \) = the maximum quantized calibrated pixel value (DN = 0) corresponding to \( L_{\text{Max}(j)} \).

\[
\rho(j) = \Pi L_{(j)} d^2 E_0(j) \cos(\theta)
\]

where, \( L_{(j)} \) = Spectral radiance at the sensor’s aperture in W/(m² sr µm), \( d \) = distance between sun and earth, \( E_0(j) \) = Solar irradiance values, \( \theta \) = Solar zenith angle.

3.3. Field work

A recognizance survey was done to work out the classification scheme (Table 2). The prime objective was to map the LULC with a special emphasis on the net sown area and cropping pattern. The database generated was part of National LULC Mapping Initiative of National Remote Sensing Center (Anonymous, 2006). Detailed ground verification was carried out for all the representative classes. Each site was recorded with GPS locations, distribution of land use land cover, field photographs and marking on satellite data and topographic sheets. The laboratory exercise were
conducted to have a uniform distribution of the GPS locations and identify the gap areas for makeover.

3.4. Decision tree approach

Decision tree classification techniques have been used successfully for a wide range of classification problems, but only recently been tested in detail by the remote sensing community (Kandrika & Roy, 2008; Pal & Mather, 2003). The classification structure defined by a decision tree is estimated from training data using a statistical procedure. See5.0, a data mining tool, was used as the rule induction engine to build decision trees. Decision trees are classification algorithms that partition a data set into homogeneous subsets. Nodes are where branches split or split the data set; terminal nodes are called leaves. See5.0 builds a tree by determining splits in the data set which minimize the entropy at a node (Quiland, 1993).

The decision on the optimal split at a given node is made according to the gain ratio criterion, which is the ratio of the gain to the split info. The gain is the change in entropy between the node and the weighted entropy across the sub-nodes stemming from the split. The split info is used to avoid bias in favor of splits with many outcomes.

The gain of a split X is:

\[
gain(X) = info(T) - \sum_{i=1}^{n} \frac{|T_i|}{|T|} info(T_i) \tag{3}
\]

where, T is the training cases at the node, Ti the training cases at the ith sub-node following split X and \(|T_i|\) gives the count. info(T) and info(Ti) are the average information of sets T and Ti, respectively (also known as the entropy), where:

\[
info(S) = -\sum_{j=1}^{K} \frac{freq(C_j, S)}{|T|} \log_2 \left( \frac{freq(C_j, S)}{|T|} \right) \tag{4}
\]

for set S and Cj identifies the jth class. The split info of a split X is given by:

\[
split \text{info}(X) = -\sum_{i=1}^{n} \frac{|T_i|}{|T|} \log_2 \left( \frac{|T_i|}{|T|} \right) \tag{5}
\]

The gain ratio measures the proportion of the information generated by the split that is beneficial to the classification. Each split is chosen so as to maximize the gain ratio and thus the information gained.

3.5. Rules generation

During training site definition phase, care has been taken to account all the temporal variabilities especially in agricultural areas. See5.0 commercially available decision tree algorithm has been used for training the temporal AWiFS datasets. Initially, digital numbers along with the land cover class details have been converted into the format suitable for See5.0 software and rule sets were generated. The output rule sets along with the trial classification results were examined for their confidence levels and accuracies. Based on these results, the training sites were modified wherever necessary till reliable training sets are obtained, and good classification accuracies were achieved. These rule sets were imported into the ERDAS imagine Knowledge Engineer module and the entire image data set has been classified. These results were re-examined visually for their class accuracies vis-a-vis temporal satellite data sets. Training sets were redefined or additional training sites were defined wherever necessary till a visually satisfactory product is attained.

The rules \(n = 232\) were generated using various cases \(n = 19,018\) randomly distributed over the different classes. Each of case was evaluated for the each class using the GPS locations and satellite data. The maximum overlaps were between the double crop and forest areas. The forests of Delhi being very thin and mostly deciduous thorny gets mixed up with agriculture areas. Table 3 provides evaluation of the training data for the rule generation.

The iterative run results in total 184 rules having varied confidence level ranging from 0.315 to 0.993. Each rule was evaluated for the level of confidence for classifying the pixel. Some of the rules were removed prior to classification owing to lesser confidence level. All the rules showing confidence greater than 0.92 (threshold) were filtered in for further processing and classification of satellite data. However for the classes having gross confidence lesser than threshold, at least one rule having maximum confidence was taken into account. Table 4 provides details of the number of rules generated and rules used for classification.

The gain ratio measures the proportion of the information generated by the split that is beneficial to the classification. Each split is chosen so as to maximize the gain ratio and thus the information gained.

3.6. Accuracy assessment

Accuracy assessment of coarse resolution land cover maps poses a great challenge to the remote sensing community. To
assess the classification accuracy, independent ground samples collected during the fieldwork, fine resolution images and other set of land cover maps have been used. The GPS points were uniformly distributed throughout the study area. A set of land cover information collected during the fieldwork of present exercise was also kept separate for accuracy assessment. The cover type information of these locations (GPS points) was compared with classified map. The field sample locations were overlaid on classified maps to assess corresponding classes. Statistically valid sampling strategy was adopted to assess commission, omission and overall accuracy (Joshi et al., 2006; Rosenfield & Fitzpatrick-Lins, 1986; Stehman, 1996).

4. Results and discussion

4.1. Land use land cover

Fig. 3 and Table 5 displays the results of decision tree classification of Delhi. Dominant LULC types are Built up land (35.78), double crop (20.57), rabi crop (10.83), kharif crop (10.31), Wasteland (4.95), Forest (4.45), plantation (5.03), current fallow (2.79), Scrubland (2.30), water bodies (1.62) and zaid crop (1.34). (Table 5).

![Fig. 3. Delhi LULC classified map using AWiFS multi-temporal dataset.](image)
The reference data pertaining to various cover classes are provided by different organizations, the data generated for various LULC classes have been compared to those data sets. To facilitate comparison, the present 11 classes were collapsed into 4 major classes namely forest, net sown area, currently fallow and others.

The forest cover was quite satisfactorily comparable with the area estimate in estimates could be due to non-inclusion of plantations/orchards/trees in the forest cover. The present study grouped plantations, orchards and tree groves into one class and was considered in forest cover. The delineation of shrub/scrub/degarded forests was not attempted considering the resolution and scale of mapping of the project. In view of this certain proportions of shrubs/scrubs/ degraded forests, which might have less than 10% of crown density, might have been included in the forest cover.

The percent coverage of current fallow (~2.79 versus 2.72) and wasteland (~4.95 versus 4.76) was very close to the figure provided by Land Use Statistics, Ministry of Agriculture (GoI, 2005). The minor differences in the estimate may be attributed to variations in the estimation techniques, resolution of satellite and variations in temporal data, which is likely to be managed with a strong ground verification and post classification enhancement. The NSA statistics was compared with the DES Statistics of 2004–2005. It may also be noted that the present study reports the areas based on the actual cover existing as on the date of satellite data used. However, DES areas refer to recorded areas, which are updated over longer time periods. Hence, the absolute comparison of the areas is not suggested.

### Table 5

<table>
<thead>
<tr>
<th>LULC classes</th>
<th>Area (ha)</th>
<th>Area (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Double crop (Db)</td>
<td>30048.21</td>
<td>20.57</td>
</tr>
<tr>
<td>Kharif crop (Kh)</td>
<td>15067.85</td>
<td>10.31</td>
</tr>
<tr>
<td>Rabid crop (Rb)</td>
<td>15828.96</td>
<td>10.83</td>
</tr>
<tr>
<td>Zaid crop (Zd)</td>
<td>1953.10</td>
<td>1.34</td>
</tr>
<tr>
<td>Current fallow (Cf)</td>
<td>4084.95</td>
<td>2.79</td>
</tr>
<tr>
<td>Scrubland (Sc)</td>
<td>3361.47</td>
<td>2.30</td>
</tr>
<tr>
<td>Forest (Fo)</td>
<td>6509.708</td>
<td>4.45</td>
</tr>
<tr>
<td>Plantation (Pl)</td>
<td>7396.81</td>
<td>5.03</td>
</tr>
<tr>
<td>Water bodies (Wb)</td>
<td>2377.71</td>
<td>1.62</td>
</tr>
<tr>
<td>Built up land (Bu)</td>
<td>52273.04</td>
<td>35.78</td>
</tr>
<tr>
<td>Wasteland (Wl)</td>
<td>7229.73</td>
<td>4.95</td>
</tr>
<tr>
<td>Total</td>
<td>146082.7</td>
<td>100</td>
</tr>
</tbody>
</table>

Table 2 also provides the description of each class mapped and its importance. It can be inferred from Table 5 that during 2004–2005, the NSA contributes the NSA. The percent coverage of current fallow (~2.79 versus 2.72) and wasteland (~4.95 versus 4.76) was very close to the figure provided by Land Use Statistics, Ministry of Agriculture (GoI, 2005). The minor differences in the estimate may be attributed to variations in the estimation techniques, resolution of satellite and variations in temporal data, which is likely to be managed with a strong ground verification and post classification enhancement. The NSA statistics was compared with the DES Statistics of 2004–2005. It may also be noted that the present study reports the areas based on the actual cover existing as on the date of satellite data used. However, DES areas refer to recorded areas, which are updated over longer time periods. Hence, the absolute comparison of the areas is not suggested.

### Table 6

<table>
<thead>
<tr>
<th>Class name</th>
<th>Reference totals</th>
<th>Classified totals</th>
<th>Number correct</th>
<th>Producers accuracy</th>
<th>Users accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Double crop (Db)</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>100.00</td>
<td>100.00</td>
</tr>
<tr>
<td>Kharif crop (Kh)</td>
<td>8</td>
<td>10</td>
<td>8</td>
<td>100.00</td>
<td>80.00</td>
</tr>
<tr>
<td>Rabid crop (Rb)</td>
<td>13</td>
<td>10</td>
<td>10</td>
<td>76.92</td>
<td>100.00</td>
</tr>
<tr>
<td>Zaid crop (Zd)</td>
<td>6</td>
<td>10</td>
<td>6</td>
<td>100.00</td>
<td>60.00</td>
</tr>
<tr>
<td>Current fallow (Cf)</td>
<td>11</td>
<td>10</td>
<td>10</td>
<td>90.91</td>
<td>100.00</td>
</tr>
<tr>
<td>Scrubland (Sc)</td>
<td>13</td>
<td>10</td>
<td>10</td>
<td>76.92</td>
<td>100.00</td>
</tr>
<tr>
<td>Forest (Fo)</td>
<td>9</td>
<td>10</td>
<td>9</td>
<td>100.00</td>
<td>90.00</td>
</tr>
<tr>
<td>Plantation (Pl)</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>100.00</td>
<td>100.00</td>
</tr>
<tr>
<td>Water bodies (Wb)</td>
<td>9</td>
<td>10</td>
<td>9</td>
<td>100.00</td>
<td>90.00</td>
</tr>
<tr>
<td>Built up land (Bu)</td>
<td>9</td>
<td>10</td>
<td>9</td>
<td>100.00</td>
<td>90.00</td>
</tr>
<tr>
<td>Wasteland (Wl)</td>
<td>12</td>
<td>10</td>
<td>10</td>
<td>83.33</td>
<td>100.00</td>
</tr>
<tr>
<td>Totals</td>
<td>110</td>
<td>110</td>
<td>101</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Overall classification accuracy = 91.81
sets was also found to play a major role in deciding the individual class accuracy. Decision tree classification technique uses a single test at each node to partition the data into homogenous groups. When multiple tests are used at each the results may improve further. To account for large landscape, attempts could be made to improve through statistical parameters.

The database is first of its kind being a product of moderate to high resolution (spatial resolution =56 m) multi-temporal satellite data with two advantages. First it is tailored to the information content of remotely sensed observations. Second, it provides relatively stable classification scheme, which can further be used to national and/or global climatic change studies. It is best suited to study the dynamics in a very small interval of time at regional scale. With this, AWiFS data can be used for the annual estimates of NSA and forest cover estimates. It shall be equally efficient and economical data at 1:250,000 scale, unlike medium resolution datasets viz., LISS III or Landsat. Even at state and/or regional level AWiFS is ideal for rapid national resource cover mapping. The advantage of sensors lies in suitable spatial (~56 m) and spectral (green, red, NIR and SWIR) resolution for vegetation delineation and temporal resolution (~5 days) for vegetation dynamics. It offers wider coverage for comparison of regional or continental level studies. Its effectiveness is in discrimination of forest types and major crops and land cover and uses. The present study recommends the use of AWiFS data for land cover and land use mapping at mesoscale for regional level assessment and monitoring.

Acknowledgments

The authors acknowledge the support from NRR initiative of ISRO/NRSC. We are also thankful to Dr. P S Roy, Deputy Director (Application Area) and other colleagues at NRSC for the support rendered during data processing and execution of the NRR first cycle project. We also thank Dr. V K Dadhwal, Dean, Indian Institute of Remote Sensing, Dehradun for constant support and useful suggestions during the course of research.

References


