# Classification of Land Cover Using Decision Trees and Multiple Reference Data Sources

E. Lieng [a,*], D. Vikhamar Schuler [b], L. Kastdalen [a], G. Fjone [c], M. Hansen [d], J.P. Bolstad [e]

[a] Hedmark University College, Evenstad, 2480 Koppang, Norway – einar.lieng@hihm.no, leif.kastdalen@hihm.no
[b] Gjøvik University College, 2802 Gjøvik, Norway. Now at: Dep. of Geosciences, University of Oslo, Norway – dagrunv@geo.uio.no
[c] Fjone, 3854 Nissedal, Norway – gunnar.fjone@forest.no
[d] South Dakota State University, Brookings, SD 57007, USA – Matthew.Hansen@sdstate.edu
[e] Directorate for Nature Management, N-7485 Trondheim, Norway – Jan-Paul.Bolstad@dirnat.no

**Abstract – Existing map databases contains valuable and accurate information that can be used as reference data for land cover classification with remotely sensed data. However, several problems occur when we try to use reference data that has been collected for a different purpose than satellite image classification. Differences in scale, legend, elapsed time since updating and ambiguities from subjective interpretation can make two maps of the same area almost incomparable. In this study, we use forest stand maps, land inventory maps and point observations from numerous sources as reference data for decision tree classification to make a land cover maps.**

**Keywords:** Land-cover maps; reference data; decision tree classification

## 1. INTRODUCTION

Norway lacks full coverage of detailed land-cover maps. The most detailed map with full national coverage exist at 1:50 000 scale, but the classes are limited to coarse categories such as forests, lakes, marches, farmland, populated areas and parks. Below the forest limit, detailed land-cover maps exist (1:10 000), containing information about both soil and forest. In mountainous areas, above the forest limit, only selected areas are mapped with respect to vegetation information. Satellite based mapping of land cover is the first milestone in the SatNat program, as a basis for vertebrate habitat mapping and GAP-analysis. The SatNat program is a collaboration between the Norwegian Directorate for Nature Management (DN) and the Norwegian Space Centre (NRS). The focus of DN's activity is area planning and management of biological values both on local and national level.

The aim of this study is: (1) to generate satellite derived land-cover maps; and (2) to explore the benefits of using existing map data bases, available from public institutions and private companies, as reference data for the image classification.

In this study, we have examined how forest stand maps, land inventory maps and point observations from numerous sources can be utilised to make land cover maps of two counties in Norway. Two different approaches are carried out to map Østfold and Sør-Trøndelag counties. The anaysis is based on decision tree classification described in Hansen et al. (2000) and Homer et al. (2002). For Østfold county the land cover map is presented, while for Sør-Trøndelag county only forest classes are presented, since the other categories still are under processing.

## 2. STUDY AREAS

Sør-Trøndelag county (270 000 inhabitants and 19 000 km$^2$ of land) has some of the most heterogeneous vegetation of Norway. From a humid coastal climate to the dry continental region in the east, there are dense spruce forests, marshes, lush deciduous forests, mountains reaching 2286 m.a.s.l. and dry inland pine forests.

Østfold county (255 000 inhabitants and 4 200 km$^2$ of land) has the same amount of forest, but far less variation. However, the small scale landscape is still quite complex. The highest point reaches 336 m.a.s.l.

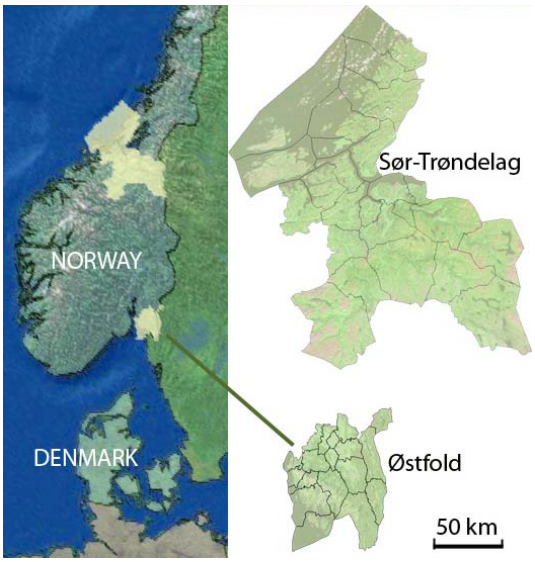



*Figure 1. The study areas Sør-Trøndelag and Østfold counties are both located in Southern Norway.*

## 3. DECISION TREE ANALYSIS

### 3.1 Overview

An overview of the classification scheme applied in this study is shown in Fig. 2. The software used to perform the decision tree analysis was S+, Cubist and See5. The satellite data and the reference data used for the analysis are described in the sections below. In general, the reference data set was split into a training data set and a test data set. The training data set was used to develop the decision trees, while the test data set was used to

* Corresponding author.

measure the accuracy of the classification on an independent data set. Some 300.000 random samples of the reference data were used to build each decision tree. During the training ambiguous decision tree nodes were adjusted to improve the quality of the classification.

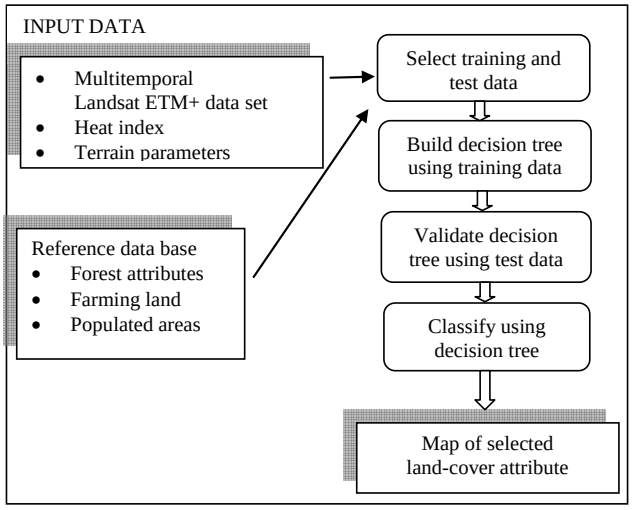



*Figure 2. Generalised scheme of the decision tree analysis.*

### 3.2 Satellite images

Primarily, we used Landsat ETM+ images for the classification. A short growing season and frequent cloud coverage over Sør-Trøndelag county generally results in few cloud-free Landsat ETM scenes per season. Therefore, some IRS-1C LISS images were added as well. To take different stages of the vegetation growing season into account, we selected three cloud-free images to represent a typical spring, summer and autumn situations. Since Østfold county has a limited extent, one multitemporal data set was sufficient to cover the entire county. In order to map Sør-Trøndelag county, which is almost five times larger than Østfold county, we needed four multitemporal data sets (Fig. 3). These data sets overlap each other to ensure comparison of the classification results. The images were co-registered with an accuracy of 0.5 pixel and aligned in a 30 meter grid. Haze, clouds, and cloud shadows were masked out manually.

For the Landsat images covering Østfold county we performed a tasseled cap transformation according to Huang et al. (2002). For the Sør-Trøndelag images, NDVI indices (Normalized difference vegetation index) were derived, and additionally an NDVI ratio of were calculated: $NDVI_{ratio} = NDVI_{summer}/NDVI_{spring}$. Ancillary data sets for the analysis included a 25 meter DEM, which was used to derive terrain parameters (slope and aspect) and a heat index which combines features of aspect and slope (Parker, 1998).

### 3.3 Reference data

Decision tree methods strongly depend on large amounts of reference data representing the classes and their spectral variability. Therefore, all available digital maps of forest stands, forest inventory and land inventory within the areas were collected and used to build and test the decision trees. These existing map data bases have been collected by various institutions, with a different purpose than being used as reference data for satellite image classification. There are three main types of data

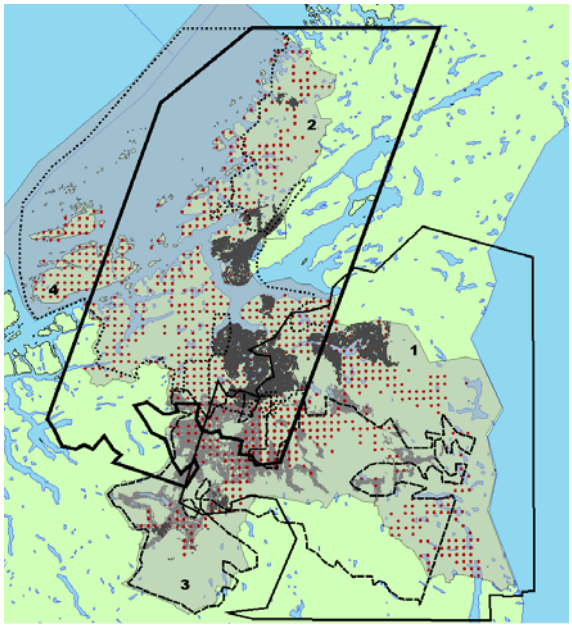



*Figure 3. Four multi-temporal dataset (1-4) from the Landsat and IRS satellites were needed to cover Sør-Trøndelag county. The reference data contains information from the national forest inventory (points), from forest stand maps (dark areas) and vegetation maps (light grey areas).*

characterising land- and forest cover available in Norway: (1) Land inventory maps; (2) Forest stand maps; and (3) National forest inventory samples.

The land inventory maps (DMK and 3Q) from the Norwegian Institute of Land Inventory are interpreted from aerial photographs. DMK covers almost the entire Norway below the forest line, while 3Q covers selected 1×1 km areas mainly within farming lands. They contain information about soil (type, quality), populated areas, vegetation type, but limited information about the forested areas (not tree species, age and forest density). In addition vegetation maps exist for a few mountain areas.

Forest stand maps are available for productive forest in most of the country, and contain information about wood volume per hectare per tree species, cutting classes (age group) and productivity classes. Priority is given to productive spruce and pine forests, and unproductive deciduous forest will be underestimated in this data set. We extracted tree species information from the wood volume.

The national forest inventory samples from the Norwegian Institute of Land Inventory are updated regularly and cover the forested areas of Norway in a 3x3 km grid. Each sample point covers 250 $m^2$. 1032 sample points are located in the forests of Sør-Trøndelag, and 1039 sample points are located in the forests of Østfold.

To use these map data bases for our analysis, extensive preprocessing was necessary. For the analysis in Østfold county, forest stand maps were converted into point data, and used as reference data for mapping forest attributes. Reference data for non-forested areas were derived from the 3Q data base. The mapped forest parameters are tree species, age groups and wood volume per hectare. An additional accuracy assessment of these

attributes was carried out using data from the national forest inventory. The accuracy was tested both by using the pixel center value and the eight nearest neighboring pixels. Success is counted if one of the eight neighbors, or the center pixel, was attributed to the estimated class. The motivation for the two tests is to account for uncertainties related to geometric coregistration of the data sets and transition zones between classes. Of the two tests, using exclusively the pixel center value is the strictest test. Open areas were mapped using reference data from the 3Q data base (farming land, meadows and rocks). Urban areas were also mapped and is described elsewhere (Vikhamar and Kastdalen, 2005). A complete description of this study is found in Vikhamar et al. (2004).

For Sør-Trøndelag county only the forest analysis is presented here. Forest stand maps were applied as reference data according to the description above. However, too small area of digital forest stand maps was available to provide the necessary amount of reference data (Fig. 3). Therefore, we increased the amount of the reference data for tree species distribution (dominance of spruce, pine or deciduous trees) by multiple regression. We used information of forest types from the widely geographically distributed land inventory maps (DMK) and indices from satellite data, DEM derived indices and distance in east and north direction to explain the percent tree cover in the the national forest inventory data set. For each of the three analysis we selected the best models based on the Akaiki Information Criterion (Akaiki, 1973). The satellite indices we used were NDVI from spring and summer, the $NDVI_{ratio}$ and the red band from both seasons. In the end, the reference data base consisted of two data sources: derived attributes from the forest stand maps and the result from the multiple regression analysis. In this way the geographic distribution of the reference data increased.

We used the result from the decision tree in the overlapping zone between handling area 1 and the other three handling areas as supplements to the other training data for these areas.

Mapped forest parameters in Sør-Trøndelag county are tree species, age, forest density and productivity classes. Forest clear cuts were identified by thresholding the NDVI and the red channel in the summer image.

## 4. CONCLUDING DISCUSSION

The final land-cover map of Østfold county is shown in fig. 4. Urban and populated areas are represented by two classes: (1) 10-50% impervious areas; and (2) more than 50% impervious areas. Open areas are represented by three classes: farming land, meadows and rocks. Three maps are produced for the forested areas (tree species, age groups and wood volume per hectare), and the tree species classes are shown on the presented map. Lakes and marches have not been mapped, but originate from the existing 1:50 000 scale land-cover maps of the Norwegian Mapping Authority. The obtained accuracy for each class is shown in Table I. The results show that particularly deciduous forest is poorly mapped, which is due to the underrepresentation of this class in the forest stand maps used as reference data. The classification accuracy of each decision tree analysis of the subsets from Sør-Trøndelag county was about 70-75%. The final tree type map only got 58%, 64% and 68% overall accuracy when compared to national forest inventory, forest stand maps and the land inventory regression map respectively (Fig. 4). Not surprisingly, as the concurrence between the reference data sets is not more than 55 to 65%.

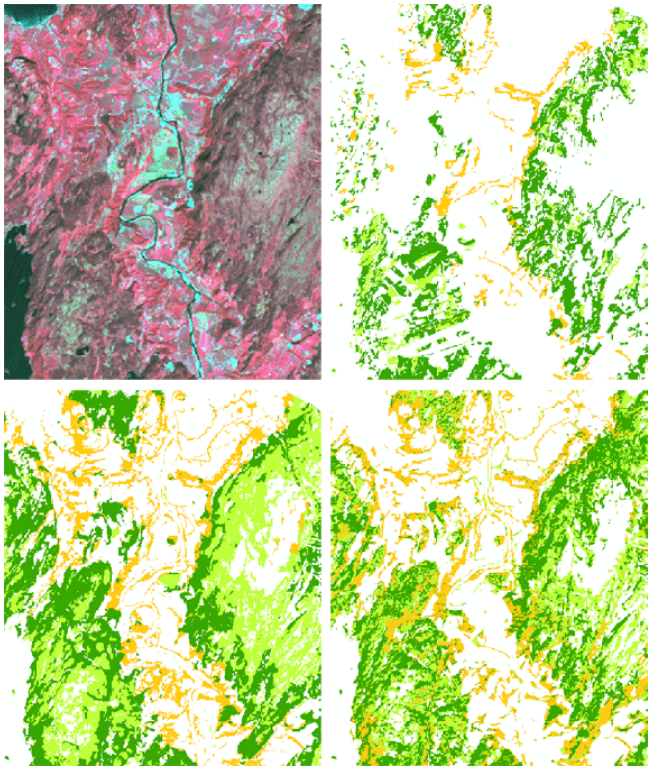


*Figure 4 Examples from the forest tree type classification, Sør-Trøndelag. Upper left: Satellite image (IRS LISS 21. July 2002), upper right: Stand data classified as tree types in the forest maps, lower left: map of training data derived from a regression (explained in the text), lower right: final classification with decision tree.*

*Table I Ranked list of the control point accuracies (%) obtained for the land-cover classes (Østfold county).*

| Land Cover Class | Center Pixel | 8 Neighboring Pixels |
|---|---|---|
| Farming fields | 94 | - |
| Low forest density | 76 | 99 |
| Old forest | 73 | 98 |
| Young forests | 54 | 90 |
| Mixed forest | 51 | 92 |
| Pine dominated | 42 | 86 |
| Medium forest density | 37 | 84 |
| Meadows | 35 | - |
| Clear cuts[a] | 34 | - |
| High forest density[a] | 21 | 61 |
| Rocks[a] | 21 | - |
| Spruce dominated[a] | 17 | 56 |
| Deciduous dominated[a] | 0 | 4 |

a. Less than 100 control points

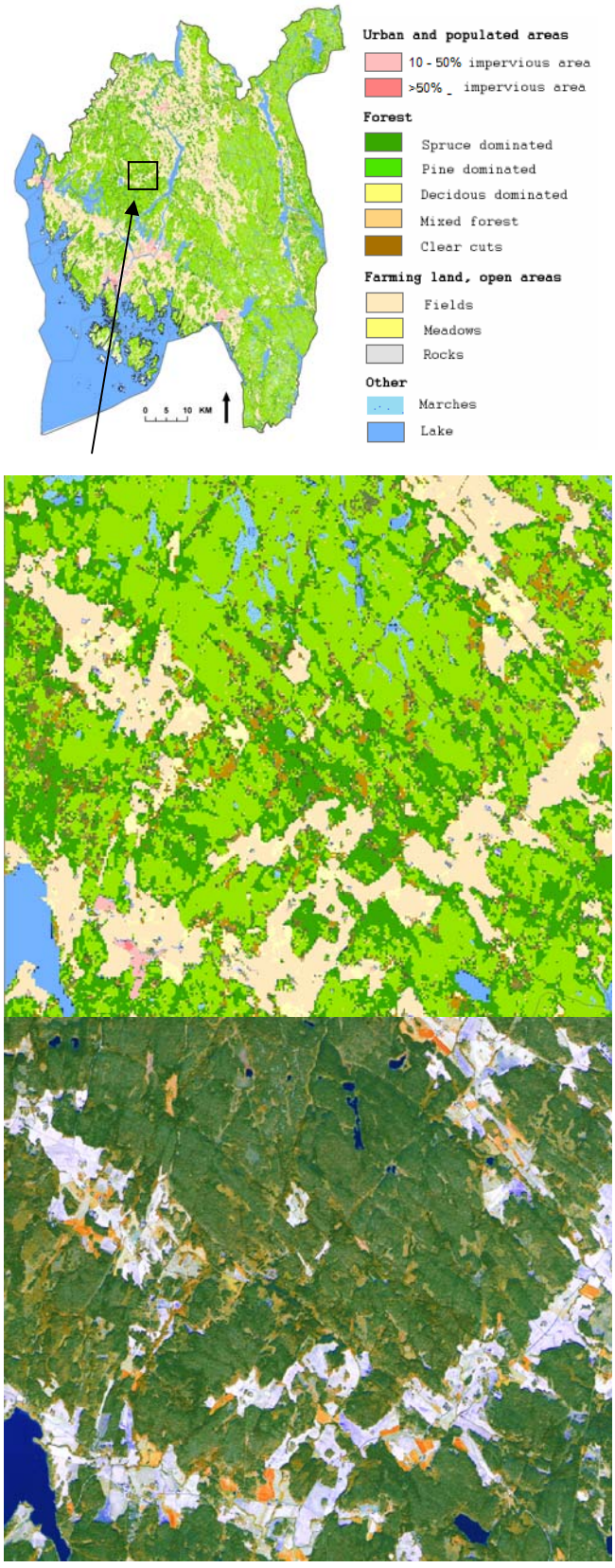



*Figure 5 Land-cover map of Østfold county. Below is a subsection of the map as well as a SPOT image to illustrate the area.*

Several factors affect the classification accuracy. Although the geometric positioning of single satellite images is approximately 0.5 pixels, the geometric errors may be larger away from the control points. Accurate coregistration of images and ancillary data sets are crucial for per-pixel classification methods. Wood volume registration errors in the forest stand data may in average reach up to 24%. The classification will never be better than the reference data used to train the algorithm. Accuracy was tested using an independent data set from the national forest inventory, where each control area constituted only 1/3 of the size of a Landsat pixel. Consequently, if the forest variables vary significantly around a control area, the overlaying Landsat pixel may be mapped to another class.

Maps showing forest parameters presented in this paper are of special interest for several applications: regional statistics and wildlife monitoring. The maps show a probability of occurrence for a certain class at a certain location. Such maps may be used to map potential habitat areas for specific animals. This requires knowledge about habitat-vegetation associations for the actual animal. As an example, the number of moose in Norway has largely increased during the last 50 years. Larger food supply is the result of an altered forestry, which increased the areas of clear cuts and young forests. Maps showing tree species, age groups and clear cuts will help predicting the food supply, and thereby the number of moose.

## 5. ACKNOWLEDGEMENTS

Forest stand data were generously made available by the forest companies Prevista AS, Foran AS, and Skogeierforeningen Nord. 3Q data, National forest inventory data, the DMK maps and the vegetation maps were provided by the Norwegian Institute of Land Inventory.

## 6. REFERENCES

Akaiki, H., 1973. Information theory and an extension of the maximum likelihood principle. In: B.N. Petran and F. Csàaki (Editors), International symposium of Information Theory. Akadèemiai Kiadi, Budapest, Hungary, pp. 267-281.

Hansen, M.C., DeFries, R.S., Townshend, J.R.G. and Sohlberg, R., 2000. Global land cover classification at 1 km spatial resolution using a classification tree approach. International Journal of Remote Sensing, 21: 1331-1364.

Homer, C., Huang, C., Yang, L. and Wylie, B., 2002. Development of a Circa 2000 Land cover Database for the United States, Proceedings of the American Society of Photogrammetry and Remote Sensing Annual Conference, Washington D.C, USA.

Huang, C., Wylie, B., Yang, L., Homer, C.G. and Zylstra, C., 2002. Derivation of a tasseled cap transformation based on Landsat 7 at-satellite reflectance. International Journal of Remote Sensing, 23(8): 1741-1748.

Parker, K.C., 1998. Environmental relationships and vegetation associates of columnar cacti in the northern Sonoran desert (Arizona, USA). Vegetatio, 78(3): 125-140.

Vikhamar, D., Fjone, G., Kastdalen, L. and Bolstad, J.P., 2004. Satellittdata til kartlegging av arealdekke. Utprøving av beslutningstremetodikk i Østfold fylke, Direktorat for Naturforvaltning, Trondheim, Norway.

Vikhamar, D. and Kastdalen, L., 2005. Impervious surface mapping in Southern Norway, 31st International Symposium on Remote Sensing of Environment, St. Petersburg, Russia.