# SPATIAL DATA MINING APPROACH TO CLASSIFY QUECUS SPECIES A CASE STUDY IN HIMALAYA

Giang Tran Thi Huong[1], Sameer Saran[2]

[1] *Cartography Dept,*
*Hanoi university of Mining and Geology, Vietnam*
*Email: giangde0912@gmail.com*
[2] *Geoinformatics Division, Indian Institute of Remote Sensing, Dehradun, India*
*Email: sameer@iirs.gov*

**KEY WORDS:** Machine Learning, Data Mining, Non parametric, Decision tree, Image classification.

**ABSTRACT:** There are many approaches have been used for image classification and decision tree is one of known practical and effective approach. However, depending on characteristic of each object, input parameters for making the tree is very important to get the best result. This study tried to find the optical parameters are used for classifying forest species by using Remote Sensing data which still one of the big problems in image classification so far. To overcome this problem decision tree is reasonable choice.

Decision tree is not new in Remote Sensing image classification. There are some researches showed that decision tree gives the better result than conventional classification approaches by using See5 data mining software combining with knowledge based classification in ERDAS imagine software. However, Classification and Regression Tree (CART) data mining software is also an advanced tool for tree-structured data analysis. And this study also shows that CART software produces results that easy to understand and easily combining with based classification ENVI imagine software help delineate the result quicker and better than to build the tree by using See5 in the case of classification Quecus species in Himalaya region.

## INTRODUCTION

Decision tree approach is a non-parametric classifier and an example of machine learning algorithm. It involves a recursive partitioning of the feature space, based on a set of rules that are learned by an analysis of the training sets. A tree structure is developed where at each branching a specific decision rule is implemented, which may involve one or more combinations of the attribute inputs. A new input vector then “travels” from the root node down through successive branches until it is placed in a specific class. The advantages of decision tree are good accuracy, short learning time and a small of memory usage.

The current study focus on collecting data for building classification tree in See5 and CART software and after that base on the output tree type Knowledge based classification is apply to classify the remote sensing image. The rest of this paper is organized as follows. In the section II we present the Study area overview. We introduce our approach to decision tree construction in section III. The comparison of efficiency is discussed in section IV.

## STUDY AREA AND DATA USED

### Study area

The study area chosen for the research is Nanital, Kumaun, Uttrakhand, India. The geographic extent of the study area is $79^0$19’ E to $79^0$40’ E and $29^0$17’ N to $29^0$30’ N latitute.The forests of Kumaun Himalaya are represented mostly by pine forests, pine mixed forests, evergreen broad leaf species (Champion and Seth, 1968). There are three main types of oak forests in Nainital: *Q. Semecarpifolia, Q. Floribunda, Q. Leuchotrichophora.*

### Data used

Two set of images were used in study. Landsat ETM required on 23th October 2002 and SRTM 90m spatial resolution.

### Elevation

Among environmental variables, elevation is the most important factor in relation to vegetation distribution. This oak is found throughout the Himalaya from Bhutan westwards; chiefly at 8,000-12,000 ft. That is the reason why elevation could help in differentiating between Quercus forests, with other types of forest. Therefore, it was decided to add as ancillary layer to discriminate these classes. Elevation map is required from SRTM image.

### Slope and Aspect map

Besides elevation, other topographic factors, such as slope and aspect, are also significant to spatial variation of plant communities, because aspect and slope also affect soil water conditions and temperature, and aspect further affects isolation in the communities (Lomolino, 2001; Krestov et al., 2006).
An aspect map shows to which side a slope is directed. The aspect of a slope can make very significant influences on its local climate (microclimate). This can have major effects on altitudinal and polar limits of tree growth and also on the distribution of vegetation that requires large quantities of moisture.
The oak forests occur in regions of heavy snowfall and at least moderate rainfall; they do not extend into the driest parts of the inner Himalaya. (Troup, 2006). This is one characteristics of this oak to decide take aspect and slope into account to be ancillary layers.
Slope and aspect map are derived from SRTM image by using ArcGIS software.

### Solar Radiation

Solar radiation is the radiant energy emitted by the sun in the form of electromagnetic wave. With different height solar flux varies and it influent to growing of oak tree. By using ArcGIS normal solar radiantion is derived.

### Texture

In pixel-based approach, each pixel is classified individually, without considering contextual information (i.e., characteristics or label assigned to neighbor pixels). Several studies have explored the potential for using these texture statistics derived from satellite imagery as input features for land cover classification. Given a feature, we might gain additional class discrimination power by considering contextual variability, in addition to the feature's ability to organize class labels based solely on its spectral values, which in turn could help in improving the accuracy of the classification.

Texture image was generated using PAN 15 m resolution using ERDAS image processing software. It is used for discriminate oak tree with other kind of tree.

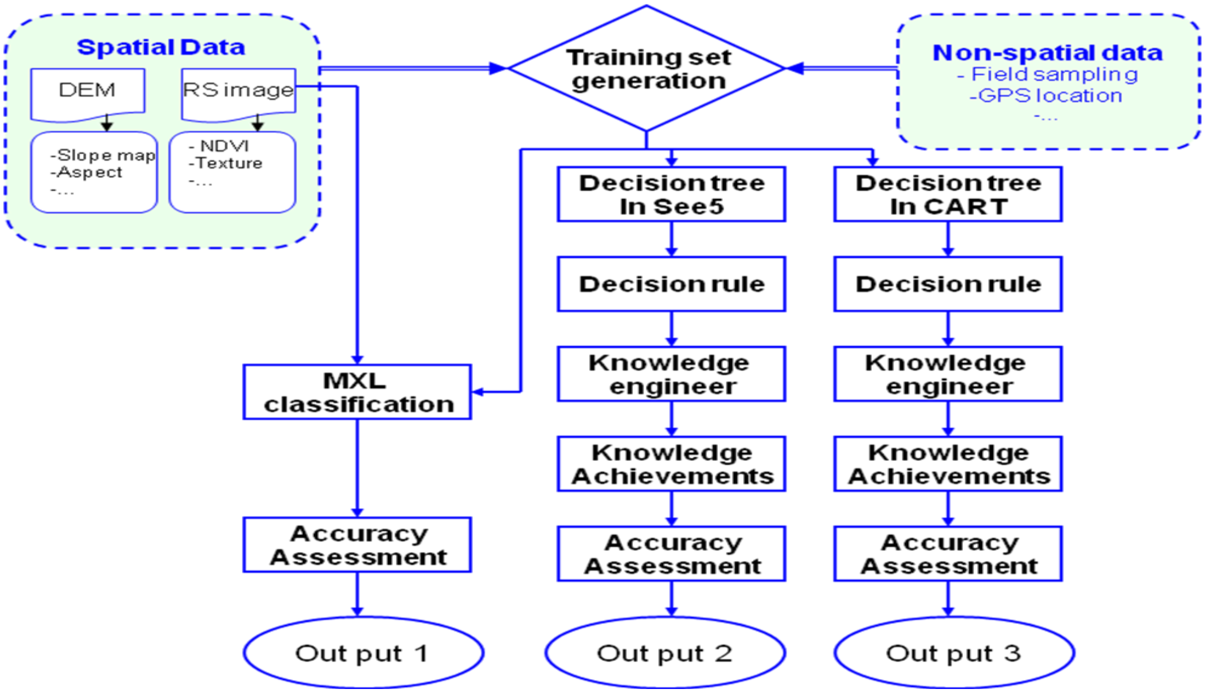
## METHODOLOGY

### Data preparation

To prepare data for decision tree layer stacking option is required. Total ten layers were taken for layer stacking operation. Four layers RED, GREEN, NIR and Thermal Infrared from Landsat image, which were at 23.5 m resolution. One texture layer generated from PAN image at 15m resolution. Slope, aspect, solar radiation and DEM are four bands extracted from SRTM image with 90m spatial resolution. The last one is NDVI layer which collected after operating NDVI calculation from ERDAS software. Re-sampling was done to bring all the ten layers to one common spatial resolution. Finally, 15m spatial resolution was chosen and all the layers were re-sampled to this pixel sizes. Nearest Neighborhood algorithm was used for re-sampling because it preserves the spectral values of the image pixels.

### Training dataset

The taken sample points are around 360 samples. After using ERDAS software to convert Pixel to ASCII file all the value of pixel in ten layers were extracted, edited, and two files were created: file.data and file.name.



## IV. RESULT AND DISCUSSION

### 1. Maximum likelihood classification

Maximum likelihood classification is done by using ERDAS image processing software. The figure below showing the classification result of Landsat ETM using Maximum likelihood classification in ERDAS image processing software:

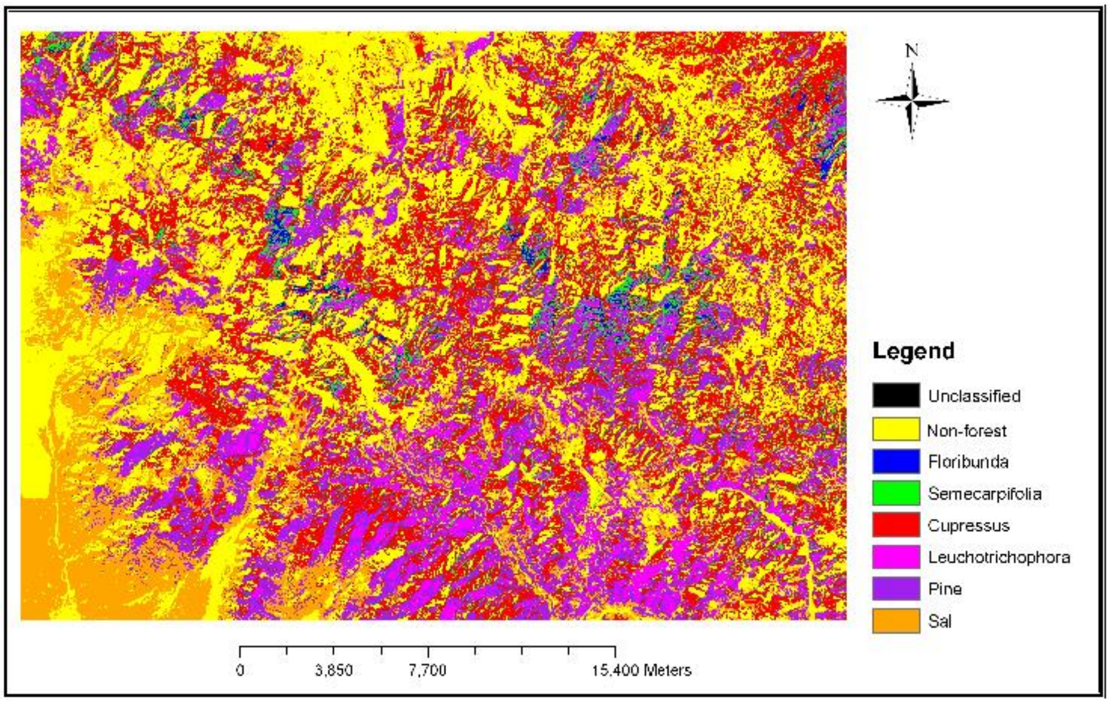


*Figure 1: Maximum likelihood classification*

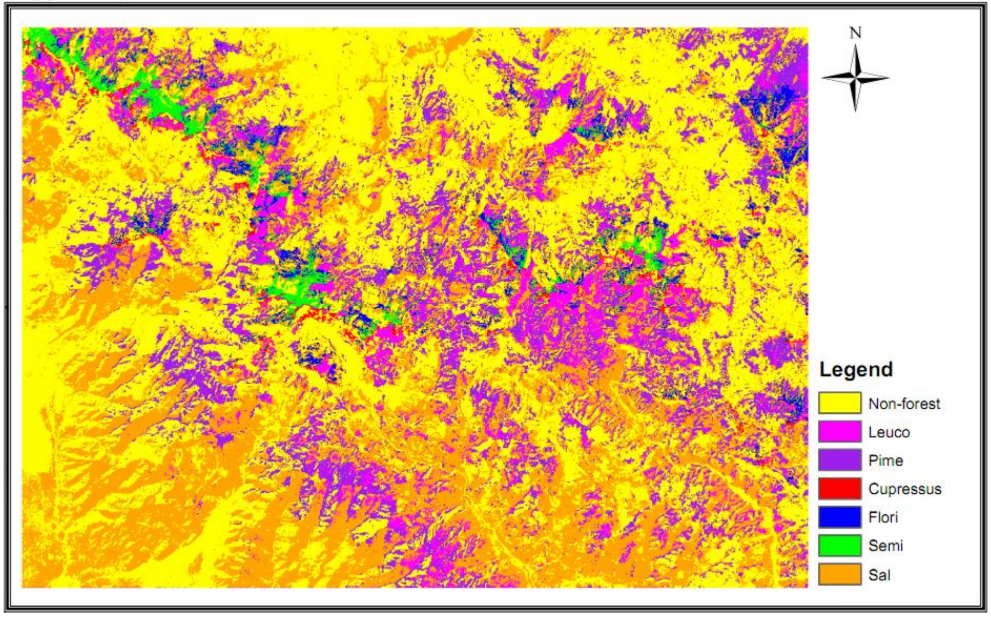
*Overall accuracy: 61.93%*
*Kappa coefficient: 0.556*

Non-forest class includes: water, urban, bare land and agriculture. The classified image gave quite good result of non-forest and forest area, because almost non-forest object have different spectral reflection with forest areas. However, water class has the reflection similar to shadow areas, so all forest area in shadow is classified into non-forest class. That is one of drawbacks of MXL classification. Moreover, Maximum likelihood's discriminating ability in forest types is not good, as well as a lot of mixed forest types. To explain for that the reason is vegetable spectral reflection for separated kinds of forest is not clear. In the result image of MXL, the area of Floribunda forest and Semecarpifolia forest is less than in fact. Most of Semecarpifolia area is misclassified with Cupressus area because of shadow.
Because just based on spectral reflection of object in the image, so Cupressus forest is mixed with the object is located at low height where pine forest locates.

**2. See5**

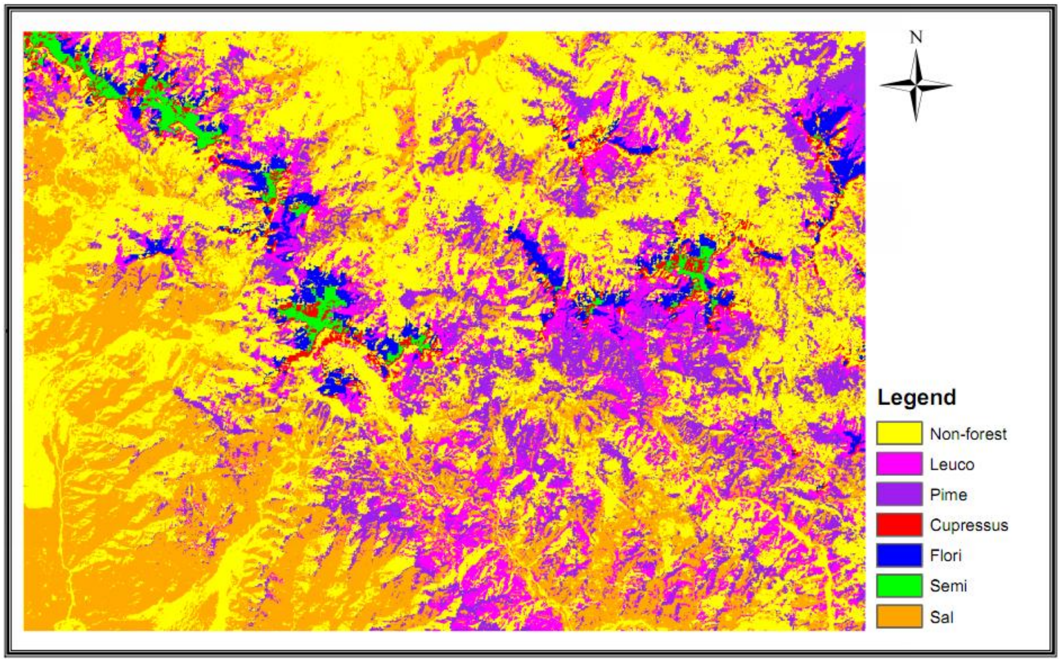
Based on knowledge of engineer about oak tree characteristic required from See5 output, the output is quite good especially with *Q. Floribunda,* and *Semecarpifolia*,  with Pine and Cupressus is also give a good result. However, the mixing between non-forest and Sal forest is occurred because of is has not much different sign between them.

*Overall accuracy: 75.57%*
*Kappa coefficient: 0.7178*



**3. CART**

CART's result is the best result. It got high accuracy compare to two last results. Because CART' algorithm chosen the best tree automatically so it not take the time for the user selects the best rules like See5. As well as, all the pixel in CART is labeled, this is difficult in See5. In See5, if you want to get the result which suitable for all pixel it will take a lot of time. Furthermore, CART deliver the error for all class so the accuracy although did not get perfect for all class but it got relative high accuracy in them.



## CONCULSION AND RECOMMENDATION

### Conclusion

Based on the characteristics of oak tree in Nainital, Uttrakhand, India, this research is found out the optimal variables which is used to be input parameters for classification forest species, they are: red band, green band, Near Infrared, Thermal band, NDVI required from Landsat ETM data, and elevation, slope, aspect, solar radiation required from SRTM image.
The output images of Maximum Likelihood, knowledge based classification running in ERDAS, and knowledge based classification running in ENVI software showed that non-parametric classifier gives the better result than parametric classifier because it used ancillary layers as well as knowledge engineer.
Decision tree is one the the non-parametric give the quick result with visual and easy to understand. Seeing to the decision rules of See5 or decision tree output of CART easy to catch the structure of the tree and know how it works.
Training set plays an important role in accuracy of the result. Pruning and Boosting may improve the accuracy of the machine learning classifiers.

### Recommendation

In the future, using decision tree in hyper-spectral may be attempted. It may be will more useful to specify more forest species.
As well as, Combine Decision tree with multi-temporal data in research will be an added advantage to the findings.

### Acknowledgement

I am very grateful to Dr Sameer Saran, SE of Geoinformatics Division of Indian Institute of Remote Sensing (IIRS), Dr. P.S Roy, Dean of IIRS and Director of CSSTEAP, Dr. S.K Saha Course Director of CSSTEAP , Mr Nguyen Ba Duy Hanoi university of Mining and Geology for their cooperation during the project work. Especially, many thanks to the **NAFOSTED** (**National Foundation for Science and Technology Development**, http://www.nafosted.gov.vn/ ) supplied financial for me to attend this conference.

## Reference

[1] Wen Zhanga, Baoxin Hua, Linhai Jinga, Murray E. Woods, and Paul Courville. Automatic Forest Species Classification using Combined LIDAR Data and Optical Imagery.

[2] M. K. Ghose , Ratika Pradhan, Sucheta Sushan Ghose Department of Computer Science and Engineering Sikkim Manipal Institute of Technology, Sikkim, INDIA. Decision Tree Classification of Remotely Sensed Satellite Data using Spectral Separability Matrix.

[3] Saran, A, Bharti,A, Sterk.G, and Raju, PLN., 2007. Comparing and Optimizing land use classification in a Himalayan area using paramteric and non parameteric approaches, *Journal of Geomatics*, Vol.1, No.1 pp. 23-32.

[4] E. Lieng, D. Vikhamar Schuler, L. Kastdalen, G. Fjone, M. Hansen, J.P. Bolstad. Classification of Land Cover Using Decision Trees and Multiple Reference Data Sources

[5] Rokach, L.; Maimon, O. (2005). "Top-down induction of decision trees classifiers-a survey". *IEEE Transactions on Systems, Man, and Cybernetics, Part C* 35: 476–487.

[6]. ^ Breiman, Leo; Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software. ISBN 978-0412048418

[8] Quinlan, J.R. (1993), C4.5: Programs for Machine Learning. California: Morgan Kaufmann Publisher, Inc.

[9]. Jensen, J.R. (1996), Introductory Digital Image Processing: A Remote Sensing Perspective. 2th Edition. Upper Saddle River: Prentice Hall.

[10]. Beena Joshi, Ashish Tewari* and Y.S. Rawat. Population Dynamics of *Quercus floribunda* Lindl. Seedlings Under Denser and Lighter Canopied Microhabitats. Nature and Science, 2009;7(1), ISSN 1545-0740

[11]. Amit Bharti. A Decision Tree Approach to Extract Knowledge for Improving Satellite Image Classification, 2004

[12] Kontoes.C.C & Rokos.D, 1996 The integration of spatial context information in an experimental knowledge based system & the supervised relaxation algorithm-two successful approaches to improving SPOT-XS classification. International Journal of Remote Sensing, 17(16), 3093-3106.

[13] Murai, H.O., S. (1997). Remote sensing image analysis using neural network & knowledge based processing. International Joural of Remote sensing, 18(4), 811-828.

[14] Vanden Berghen Frank. 7-7-2003. Classication Trees: C4.5

[15] S. Rasoul Safavian and David Landgrebe. A Survey of Decision Tree Classifier Methodology

[16] Matthew N. Anyanwu and Sajjan G. Shiva. Comparative Analysis of Serial Decision Tree Classification Algorithms.

[17] Ravi Kothari and Ming Dong. Decision tree for classification: A review and some new results.

[18] Tom M. Mitchell (1997) Machine Learning p.2.

[19] Cormen, T. H., Leiserson, C. E. and Rivest, R. L. 1989. Introduction to algorithms, MIT Press, Cambridge, MA.