ISSN: 0974-2115
www.jchps.com Journal of Chemical and Pharmaceutical Sciences

Crop yield estimation using Isodata clustering algorithm on EO-1 Hyperion data a case study of Coconut crop, Kozhikode, Kerala

Meera Mohan K[1*], Shanmugha Sundaram G.A[2]
[1]Centre for Excellence in Computational Engineering and Networking, ASE- Coimbatore,
Amrita Vishwa Vidyapeetham University, Coimbatore-641112, TN, India
[2]Department of Electronics and Communications Engineering, ASE-Coimbatore
Amrita Vishwa Vidyapeetham University, Coimbatore-641112, TN, India
***Corresponding author: E-Mail: chinnumeera@gmail.com**

ABSTRACT

Remote sensing is one of the major techniques employed for studying agricultural patterns which tend to be highly dynamic these days. This paper delves into the possibility of finding the yield of coconut crop with hyper spectral remote sensing technology. Hyper spectral images with the advantage of having hundreds of contiguous spectral bands make crop yield monitoring a reality. Yield monitoring and crop growth assessment are very important to be carried out both at the state and national level of a country to estimate the production. In this work coconut yield is estimated by using hyper spectral images captured by the Hyperion sensor deployed on the EO-1 satellite of NASA. Unsupervised classification techniques have been used to identify the different classes present in the dataset. The process of identifying different classes extends from locating the crop under study among the different features on ground to identifying variations in the crop according to the health of the plant, pest infestation, and various other factors. The outcome of the work is evaluated in terms of the spectral reflectance of different classes. The canopy reflectance has a direct implication on crop chlorophyll and thereby their health throwing insight into the yield of the crop. Results obtained are compared with real time reflectance values measured using spectro radiometer and it is found that the hyper spectral image processing is performing on par with the hardware measurements. The health condition which is derived from these reflectance curves gives information about the yield (in lakh nuts) per acre for each class of the crop.

KEY WORDS: hyper spectral, hyperion, yield, unsupervised, classification, reflectance.

1. INTRODUCTION

In the present context, Precision farming is gaining lot of importance owing to the development in key technologies like Remote Sensing (RS), Geographic Information System (GIS) and Global Positioning System (GPS). Precision farming helps the farmers to have prior knowledge about the yield with the available raw inputs thereby having the fair chance of raising the productivity. Also surveying the yield and examining the crop growth stage are very important steps in measuring the annual productivity especially when the crop plays a vital role in the country's income and food habits. Thus, the gravity of remote sensing application in agriculture is rising. With the advent of hyper spectral imageries, information available to study is enormous. Hyper spectral Images have hundreds of continuous bands increasing the spatial and spectral resolution of the data. Recent literature laid out that the numerous bands will provide better knowledge about the physical and biological features on the ground when compared with the few banded multispectral images. Each pixel in the hyper spectral image is an array of pixels and they give a continuous spectrum, known as the spectral reflectance of the feature, which is studied to infer information about the area of interest.

Hyperion, a hyper spectral sensor mounted on the Earth Observing-1 (EO-1) satellite of NASA, provides the image with 242 spectral bands having a spectral resolution of 10nm and spatial resolution of 30m. This sensor has spectral coverage from 355nm-2500nm covering the near Infrared region, visible region, SWIR region of the EM spectrum. Narrow strip, swath width of 7.5km and swath length of 42km or 185km, of data are sourced from the earth explorer database of USGS (United States Geological Survey). Before any image processing is done for analysis purpose, the image has to be pre-processed to eliminate noise, atmospheric disturbance and other undesirable factors. Hyper spectral data always come as a cube where hundreds of bands are stacked thus penetrates more into features on ground giving better details. Multiple factors affecting the growth of a crop will have direct effect on its chlorophyll content which will be evident in its spectral response. Owing to the continuous band spectrum obtainable from the Hyperion images even the difference between mature and immature plant, pest infested plant, the degree of infestation can be identified and this part of the work plays a crucial role in understanding crop yield.

Crop identification includes classification of the enormous data which may contain several different features like vegetation, water bodies, barren land, rocky areas, and urban areas and so on. For identifying the crop of interest machine learning techniques are used. From the Hyperion data, area under coconut cultivation have to be identified and for this purpose supervised classification technique of SVM is used. The classes can be fixed by photo interpretation or prior knowledge about the site as mentioned in the work. Using this learning method when the crop is identified, the variations among the crop depending on their health conditions are to be found. For this purpose

unsupervised classification is used. Unsupervised classifier by definition classifies image pixels that have similar spectral characteristics into one class. Clustering algorithm forms the basis of the classification and two among them are K-means and Isodata algorithm. The resultant of unsupervised classification will be several different classes all representing varieties of coconut. The spectral responses of these classes are studied to understand the yield of the area considered. The classification techniques and the study of spectral signatures are carried out in an image processing environment called ENVI 4.7. The paper is arranged in sections, section II explains the study area, Section III deals with the methodology, followed by section IV discussing the results and finally with section V the paper concludes.

Study area: The study area is chosen as the Kunnamangalam region in Kozhikode district, Kerala, India having the center latitude as 10°42'00.86"N and the center longitude as 75°41'26.10"E. The Hyperion sensor has 7.7km swath width and running for a swath length of 42km or 185km. The data considered is shown in the Fig. 1. Kozhikode stands first in coconut production in Kerala and geographically the area is rich in different kinds of vegetation. Therefore, the area is well suited for analysis of coconut cultivation. The dataset has minimum cloud cover of 0-9% and the entity id of the data EO1H1450522007054110PZ_1R gives information about the dataset. ENVI can plot precise pixel-wise spectrum of the image and hence, the Hyperion image can be used to identify features and their classification.

2. METHODOLOGY

After selection of the Hyperion dataset, it should be preprocessed because it is necessary that the data should meet all the radiometric properties. Following the pre-processing supervised classification is performed and then the unsupervised classification. The spectral response of these features is further analyzed.

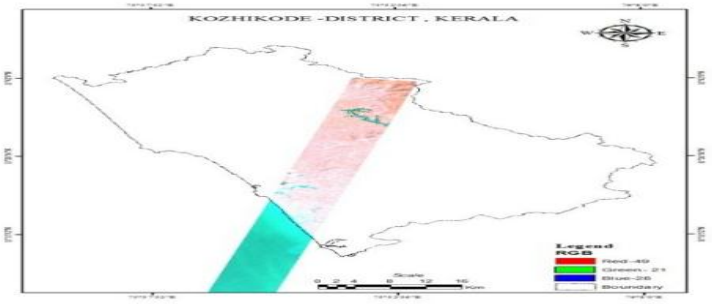

Fig. 1. Study area

2.1. Data Pre-processing: The data was acquired over the study area on 23 February, 2007 and it was downloaded free of cost from USGS archive. Hyperion images are available in level 0R (L0R), level 1R (L1R), level 1Gs (L1Gs) and level 1Gst (L1 Gst) formats, in this study the image is obtained in L1R format which is the HDF (High Definition Format) version. These images are radio metrically corrected but needed to be geometrically and atmospherically corrected and further noise eliminated. The data is pre-processed in ENVI and hdr format data is converted into ENVI file formats. Subsequently, the non-calibrated bands which are 1-7, 58-76, 77-78, 225-242 are eliminated and bands which shows undesirable effects like water absorption (bands: 120-132, 165-182, 185-187, 221-224), low SNR value (band: 77 and 78) and bands which show stripping effects (bands: 8-9, 56-57, 79-82, 97-99, 133-134, 152-153, 188, 213-216, 219-220)[5] are also eliminated.

After layer stacking the bands, georeferencing was applied to the image using GCPs (Ground Control Points) and the resultant is a geometrically corrected image. Subsequently, the image is atmospherically corrected. Out of the many algorithms provided by ENVI, dark object subtraction is chosen and the algorithm is based on selecting the darkest pixel in the image and subtracting its reflectance value from the rest of the pixels (since reflectance at the darkest pixel is assumed to be only due to the atmosphere. The result appears with more clarity than the input image (see Fig. 2a and 2b). Finally, the noise removal is done using MNF (minimum noise fraction) transform, which removes the usual sensor noise during image analysis. The output image of the last process is used for further processing (Fig. 2c).

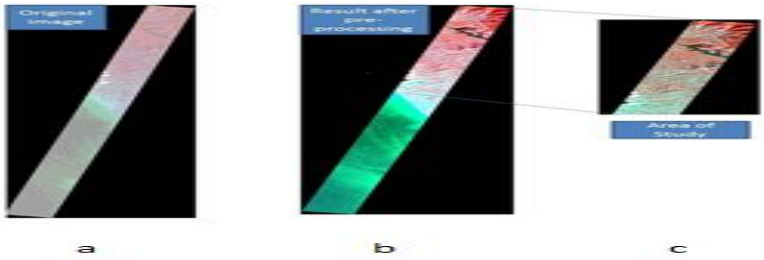

Fig. 2. a) Downloaded data set b) Atmospherically corrected image c) Subsetted AOI

2.2. Supervised Classification: Support Vector Machines (SVM) is a powerful statistical learning algorithm for performing regression analysis and classification. SVM algorithm formulates a hyper plane which is the separation surface between the classes. A vector characterizes every training class and the SVM classifier calculates an ideal hyper plane based on which all the pixels in the dataset are classified. SVM incorporates kernel based classification techniques when the data is non-linear as shown in Fig. 3 and here, the pixel are highly random in distribution and thus it is the data is very non-linear.

The kernel is non-linear which helps in identifying a non-linear hyper plane and the mapping of data into a 3-D space is shown in Fig. 4. The type of kernel chosen is Radial Basis Function (RBF) which performs inner product between pairs of classes for classification. Radial basis function is given by the formula (1)

$$k(x, y) = \exp(-\sigma \|x - y\|^2) \quad (1),$$

Where σ represents a positive parameter, that controls the radius.

When $\exp(-\sigma \|x - y\|^2)$ [8] is expanded, the result is as given in (2) which is as shown in the works,

$$\exp(-\sigma \|x - y\|^2) = \exp(-\sigma \|x\|^2) \exp(-\sigma \|y\|^2) \exp(2\sigma x^T y) \quad (2)$$

The expansion for the last term is as given by (3),

$$\exp(2\sigma x^T y) = 1 + 2\sigma x^T y + \frac{(2\sigma)^2}{2!}(x^T y)^2 + \frac{(2\sigma)^3}{3!}(x^T y)^3 + ... \quad (3)$$

Where, $\exp(2\sigma x^T y)$ is the sum of infinite polynomials [9]. Therefore, here kernel maps the function to a space in infinite dimension. The product of two kernels is another kernel therefore, the product of $\exp(-\sigma \|x\|^2)$ and $\exp(-\sigma \|y\|^2)$ is also a kernel. Thus,

$k(x, y) = \exp(-\sigma \|x - y\|^2)$ is also a kernel. With high dimensionality data like Hyperion data, RBF provides better classification results.

Before supervised classification begins, it is very essential to define the classes. These fundamentals can be formulated by photo analysis, interpretation of ALI data which has a spatial resolution of 10m, and previous understanding of the site. In ENVI, the training sets are chosen by sampling the image and storing them as region of interest (ROIs). These ROIs are collected with great care and limited to region where class variation is constant. A total of 7580 pixels are taken for training as given in Table I. With these classes SVM is applied to the dataset in ENVI.

ENVI provides SVM option in its classification tools. After defining the ROIs, the input images has to be loaded and the parameters like gamma in kernel function, pyramid levels and penalty parameters should also be defined before performing the process. Gamma is normally set as the inverse of number of spectral bands (here number of spectral bands after pre-processing is 132 and 1/132 = 0.007), pyramid level is set to its minimum value (that is, 0) and penalty parameter to 120. Finally, classification probability is set to 0.05 causing almost all pixels to be classified.

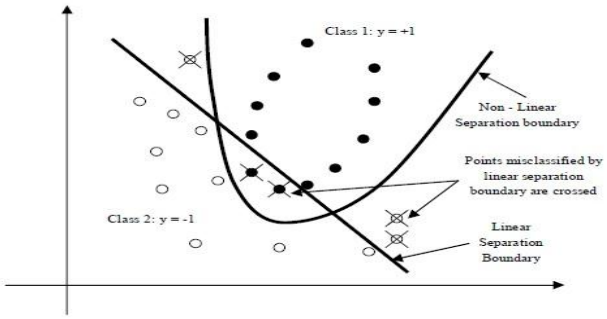

Fig.3.Linear and non-linear hyper plane

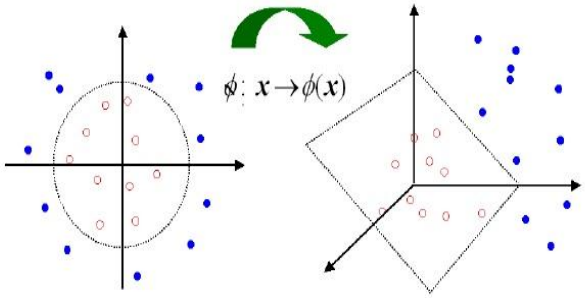

Fig. 4. Non-linear data is mapped onto a 3-D space

Table.1.Key for classification of the study area

ROI	Id	Color	Pixels	Class Description
Rocky Vegetation	1	Red	790	Sparsely vegetated area
Rock	2	Magenta	796	Area of Rocks
Water	3	Blue	217	Area with water bodies
Barren Land	4	Yellow	728	Area left with no vegetation
Coconut	5	Green	2087	Coconut cultivation
Urban	6	Cyan	2382	Settlement area
Cultivated Area	7	Maroon	580	Area under cultivation

After specifying the parameters and selecting output location clicking ok would start the computation and finally the result will obtained as shown in Fig. 6.

2.3. Unsupervised Classification: Unsupervised classification developed in ENVI classifies the vegetation area identified in the previous step (SVM method). Unlike supervised classification the data is unlabeled here. Classes are not defined in prior but the pixels are separated into classes called clusters based on their spectral reflectance properties. Unsupervised classification works on the basis of clustering algorithm where pixels are grouped automatically into classes. When SVM fixes a hyper plane for separation, the unsupervised classification strives to find an unknown structure within the pixels. K-means and Isodata are two commonly used classification algorithms. In both cases, clustering is said to have optimal performance when the clusters built represents the variation in the scene to the maximum.

Isodata is the acronym for Iterative Self-organizing Data Analysis Technique and as the name suggests it divides and combines clusters through many iterations until a predefined limit is reached. With the unlabeled classes and threshold values defined, Isodata calculates the mean of the data and the repetitively clusters the pixels until the iteration stops. Mean is recalculated in every iteration and every time pixels get reclassified. Isodata starts operating with an arbitrarily placed cluster center and a pixel set as its initial mean position. Now the rest of the pixels are clustered using minimum distance analysis which is repeated throughout the iterations. Cluster performance is greatly dependent on its threshold parameters, that is, the standard deviation calculated within a cluster if it varies from the defined maximum class standard deviation, the clusters splits and it merges when the minimum distance less than the defined value is. As the iteration proceeds clusters get eliminated if number of pixels falls below the specified value. Also the range of expected number of classes and number of pairs to be merged in the course of operation are pre-defined as shown in Fig. 5.

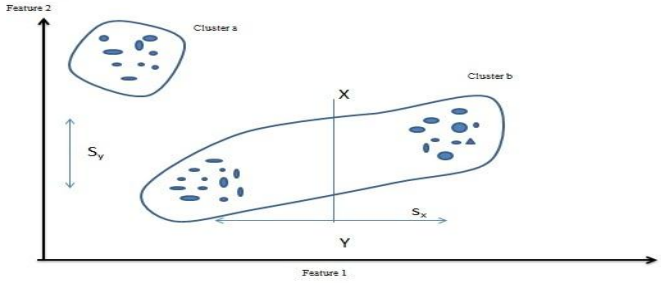

Fig.5.Isodata working algorithm

Mathematically, the objective of Isodata is to minimize the Mean Squared Error (MSE) which measure inside cluster variability.

$$MSE = \frac{\sum_{\forall \chi}\left[x - C(x)\right]^2}{(N-c)b} = \frac{SS_{distances}}{(N-c)b} \qquad (4)$$

$$SS_{distances} = \sum_{\forall \chi}\left[x - C(x)\right]^2 \qquad (5)$$

Where C(x) is the cluster mean that pixel *x* belongs to, N is the number of pixels, c indicates the number of clusters, b is the number of spectral bands.

In ENVI, Isodata clustering is performed by selecting unsupervised classification from the classification menu.

<< Set the minimum number of classes after the process to be in the range 5 to 10.

<< Maximum iterations, required for the entire classification, is given as 100.

<< Each class is formed such that they have minimum of 1000 pixels in it.

<< The maximum standard deviation is given as 1 and minimum distance between classes is given as 5.

<< The merge pairs, i.e., the number of classes that can merge if the distance between them is lower than the given value, is defined here as 2.

<< Finally a change threshold of 5% is given.

<< Output location is fixed and the Isodata clustering is computed to obtain a result as shown in Fig. 7.

3. RESULTS AND DISCUSSION

The result of SVM which produced the land cover mapping of the study area is shown in Fig. 2b and Fig. 2c shows the different classes. Through visual inspection seven predominant classes where identified and coconut plantation is one among them which spans the area in green color. For understanding the yield of the crop it takes more than just finding extend of cultivation. A detailed indexing and assessment of infested plants are necessary to evaluate the crop health. SVM classification limits mapping the features with consistent spectral response at a coarser level. Variation in coconut crop can only be understood by further classifying the identified crops extend.

The high resolution satellite imagery has undergone unsupervised classification which generates classes based on the spectral properties of the ground features. Thus, the classified image is itself expected to act as the

identification key. Unsupervised classification acts as an important tool for analysing digital images. It is very important to choose the correct classification algorithm eventually producing the final product for crop analysis. The Isodata classification algorithm has been applied to a region of 2648.843 acres of land which is identified to be under coconut cultivation. To find the variation in the crop, region is classified into six classes depending on the spectral properties difference. This can be seen in Fig. 7. The classes into which pixels got classified are shown in different colors in Fig. 7b and Fig. 7c. Class 3 and 6 from their spectral reflectance prove that they are settlement and cultivated area. The four remaining classes show variations in the spectral response and thus showing crop variations. These classes are given in Cyan, Red, Green and Yellow respectively.

3.1. Study of Spectral Properties: Fig. 8 shows the plot depicting difference in the spectral properties of these crops. Hyperion band numbers and Digital number (DN) values are marked along the x and y direction of the plot respectively. The DN values are the uncalibrated reflectance values at every pixel. From the previous works on coconut health using hardware, it has been found that the response curve maxima falls at wavelengths 485nm, 830nm, 1650nm and the curve is most responsive in the range of 700-990nm. Similar result is found in Fig. 8, peaks are found at band numbers 14, 48 and in the range of 50-80. Wavelength corresponding to the range of band numbers from 10 to 120 for Hyperion sensor is given in the Table. II. The spectral response curve of a healthy plant is said to be the plot with maximum peak then comes the slightly low yield, followed by moderate yield plants and then finally the very low yielding plants. To determine the healthy out of these, the spectral responses are compared with real time values measured using USB-4000 spectro radiometer. According to a healthy coconut leaf is said to show maximum reflectance in the range of 700-920 nm which is shown in the Fig. 9. Similar results are observed in the resultant graph of Isodata clustering in Fig. 8.The difference between healthy and slightly low yield is very less and they almost coincides in the NIR region. Moderate yield and very low yield comes further below showing remarkable difference in their characteristics.

3.2. Histogram analysis and yield per acre: The yield also depends on finding the area these classes are covering in the study site. In terms of acres of land if the crop extent is found for each class (namely, healthy, slightly low, moderate, very low) then exact idea of yield per acre can be deduced. The Class Statistics information is available in the following Table 2 which gives information about the number of pixels covered by each class and the corresponding area in acres.

In ENVI, the class statistics can be calculated as a part of post classification operation.

- Firstly, select the classification file (result of Isodata classification) and the input file in which the classification details are available (input to Isodata classification).
- Now select the classes whose statistics are to be calculated. Here, high yield, slightly low, moderate yield and very low yield are the selected classes.
- Statistics calculation includes basic statistics (number of pixels in a class, area covered in acres and percentage of coverage), histogram calculation giving the extent of different classes in the various bands and lastly the covariance and mean.
- The statistics result can be outputted either as a text report or graphically on the screen.
- Table.2 shows the report of basic statistics and Fig. 6 and Fig. 7 shows the graphical result.

Knowing the spatial resolution (30m), the area of a single pixel is 30m×30m. Then the area of a class can be found by simply multiplying the pixel number in the class and the area of a single pixel (30m×30m). For example, high yield variety has 986 pixels with each having area of 900 m^2, now the area of class is 887400 m^2 (900×986).

1 acre is 0.00024711 m^2 therefore area comes to 219 acres (which matches with the value in Table.2).The Table.2 shows that the Moderate yield coconut plant has the highest extend of 919.2992 acres. This is also shown in Fig. 10, which is the histogram representation of all classes, showing that the moderately yielding coconut plants cover is the highest as the green color plot is having the maximum frequency in all the three bands (49, 21 and 28) considered. The second common type of coconut is the slightly low yielding then very low yielding variety follows and finally the healthiest.

Table.2.Class statistics summary

Class name	No. of pixels in the class	% of area covered	Area in acres	Production in lakh nuts in different classes
High yield	986	8.046	218.9971	5.97432
Slightly low yield	3966	32.766	880.8748	24.03026
Moderate yield	4139	34.195	919.2992	25.07848
Very low yield	1383	11.426	307.1734	8.37969

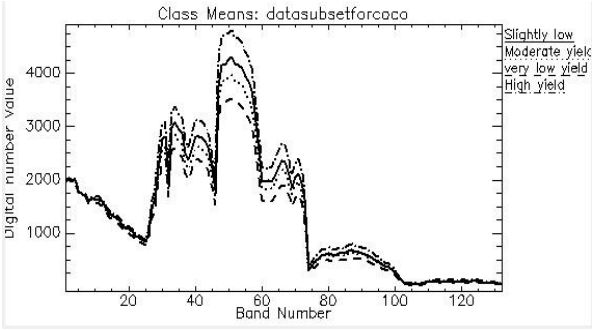

Fig.6.Result obtained from unsupervised classification

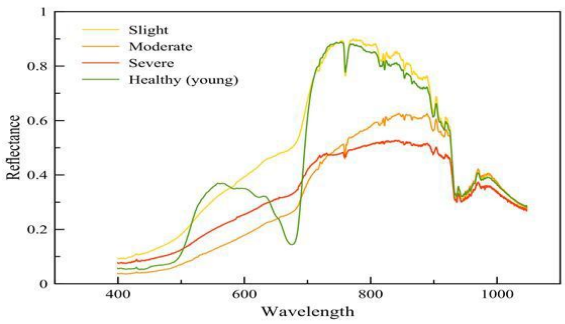

Fig. 7. Plot as measured by hardware

The normal spacing between coconut plants when planting them is 7.6m. In an acre (which is 4046.856m^2) on an average 65 coconut trees can be planted and a good yielding tall tree can produce upto 100-130 nuts per year. But considering the mixed health conditions of the crop, according to, Coconut Development Board (CDB) Kerala and Tamilnadu Agricultural University, one tree is said to produce 70-80 nuts per year on an average. Thus according to statistics given by CDB on the annual production at Kunnamangalam block of Kozhikode district the production is 2728 nuts/acre. And for the total of 2326.3445 acres 63.46268 lakh nuts are produced in one year.

Out of these total amount of nuts, maximum yield is reported to be coming from the moderate yielding class (25.07848 lakh nuts) then the slightly low yield variety (24.03026 lakh nuts) followed by very low yielding variety (8.37969) and finally the high yield class (5.97432)see Table III. The result is subjected to conditions like time of planting the seed, time of the year of harvesting and the health conditions of the plant.

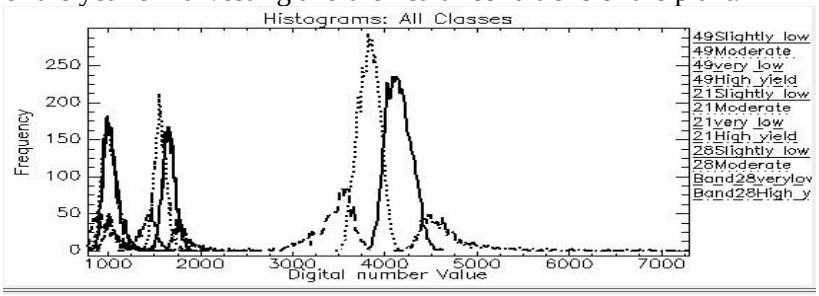

Fig.8. Histogram of all the class in R (band 49), G (Band 21) and B (band 28)

4. CONCLUSION

SVM and Isodata clustering are being used for classification of the Hyperion data. SVM is a boundary based method and it does not assumes on the distributional pattern of the data and therefore provides great results when applied to high dimension data with low sampling rate. And also with the introduction of kernel based SVMs more non-linear high dimensional data are handled with high accuracy. An unsupervised learning helps in identifying more details of a feature as it learns deeper than supervised methods which are binary classifiers. Therefore, here Isodata is used for studying the plant yield based on their health conditions. Both SVM and Isodata are hard classification techniques, i.e., they classify at pixel levels. These hard classifiers prove to a better classifiers than soft classifiers (sub-pixel classifiers) when it is hard to identify the probability function of the class and classification relies on margin based methods. In this case Hyperion data of Kozhikode chosen is a real time data with no ground truth so SVM and Isodata provided a better insight into its feature details. With the abundant information in the Hyperion data, it was able to understand the different levels of yield in coconut crop. Also the potential of studying about plants and their health in detail has been exploited. In this work the spectral response helps in finding plant health as it is directly proportional to the chlorophyll content. The result also includes area of these classes (differing in their yield levels) in acres. The yield of coconut crop in lakh nuts per acre is calculated individually for these classes which can be totaled to get the consolidated yield in lakh nuts for the area of study.

For a country like India where agricultural income plays a vital role in its economy it is always necessary to calculate the crop yield. With a single hyper spectral image if the yield over acres of land can be calculated then it helps to reduce the labor involved and time for estimation when compared with other counterparts. Coconut is one major crop that contributes to the diet and economy of most of the Asian countries. In India alone 92% of coconut production comes from south India in which Kerala produces the maximum. Therefore, calculating the yield of coconut over a large area in Kerala would definitely help in understanding the pattern of the crop and will also help the authorities to seek for the real cause of decrease in the yield. If multi-temporal data of the same site is available this work could be extended in calculating the yield and growth extend every six months thereby analyzing the growth rate and factors influencing the growth.

5. ACKNOWLEDGMENT

The authors express their deep sense of gratitude to the Hyperion data provider USGS, Earthexplorer.com. And also we are thankful to the Department of CEN (Center for Excellence in Computational Engineering and Networking), Amrita Vishwa Vidyapeetham.

REFERENCES

Beck, Richard, EO-1 user guide v. 2.3, Department of Geography University of Cincinnati, 2003.

Brisco B, Precision agriculture and the role of remote sensing: a review, Canadian Journal of Remote Sensing, 24 (3), 1998, 315-327.

Gualtieri J Anthony, and Robert F Cromp, Support vector machines for hyper spectral remote sensing classification, The 27th AIPR Workshop: Advances in Computer-Assisted Recognition, International Society for Optics and Photonics, 1999.

Kolluru. Pavan Kumar, SVM Based dimensionlity reduction and classification of hyper spectral data, The Netherlands, 2013.

Liu, Yufeng, Hao Helen Zhang, and Yichao Wu, Hard or soft classification? Large-margin unified machines, Journal of the American Statistical Association, 106 (493), 2011, 166-177.

Luo, Bin, Crop yield estimation based on unsupervised linear un mixing of multi date hyper spectral imagery, Geoscience and Remote Sensing, IEEE Transactions, 51 (1), 2013, 162-173.

Moses, Wesley Jeremiah, Atmospheric Correction of Hyper spectral Data–Execution of and Comparison between Flaash and Tafkaa_6S, Diss. Cornell University, 2005.

Petropoulos, George P, Kostas Arvanitis, and Nick Sigrimis. Hyperion hyper spectral imagery analysis combined with machine learning classifiers for land use/cover mapping, Expert systems with Applications, 39(3), 2012, 3800-3809.

Soman K.P, Loganathan R, and Ajay V, Machine Learning with SVM and other Kernel methods, PHI Learning Pvt. Ltd., 2009.

Thenkabail, Prasad S, Ronald B Smith, and Eddy De Pauw, Hyper spectral vegetation indices and their relationships with agricultural crop characteristics, Remote sensing of Environment, 71 (2), 2000, 158-182.