Variation Analysis Using Regression for Predicting Yield

Patel Shivang, Dave Maharsh, Patel Rahul, Prof. Vricha Chavan, Prof. R N Awale
Department of Computer Engineering
K J Somaiya Institute of Engineering and Information Technology, Sion, Mumbai
p.shivang97@gmail.com, maharsh.dave@gmail.com, ranchodp0497@gmail.com, vchavan@somaiya.edu, rnawale@vjti.ac.in

Abstract: **Agriculture being an important aspect of any nation can largely benefit if the pattern of growth of a crop is identified. In this study, the relation between GNDVI (Green Normalized Difference Vegetation Index) and yield is studied to predict yield based on satellite imagery obtained from landsat8 program. To identify these patterns, machine learning can provide non-biased outputs which in turn provides better approximation of yield. GNDVI values can be used to derive leaf area index and "greenness" which indicates health of a plant. If leaf area index is high it denotes that there is more amount of vegetation present in an area. GNDVI values are directly proportional to health of a plant. In this approach Regression analysis is used in order to study the relation between GNDVI values and crop yield.**

Keywords: **remote sensing; regression; machine learning, GNDVI, yield;**

INTRODUCTION

In today's growing world more than twenty-nine million hectares of agricultural land is devoted globally to growing sugarcane, producing approximately 1.8 trillion tons of raw sugar each year. Accurate and timely prediction of yield offers the global sugar industry improved efficiency and profitability by supporting decision making processes such as crop harvesting scheduling, marketing, milling and forward selling strategies. The importance of remote sensing in today's world is far more than comprehensible to human world. The agricultural world finds many unimaginative roots to make increments in the yield of crop. Remote sensing helps the farmers in measuring the yield of the crop, monitor health of the crop and what measures are to be taken to increase the same. Regression analysis models the relationship between a response variable and one or more predictor variables [1].

Spectral vegetation indices that are based on green and near infrared reflections have the high correlation with leaf area index and canopy cover (Broge and Leblanc, 2000). However, in sparse vegetated areas, the reflection of soil and sand are much higher than reflection of vegetation and so detection of vegetation cover reflection is difficult. The GNDVI values measure the "greenness" of plant which in turn provides health of the crop. "NDVI is often used worldwide to monitor drought, monitor and predict agricultural production, assist in predicting hazardous fire zones, and map desert encroachment. The NDVI is preferred for global vegetation monitoring because it helps to compensate for changing illumination conditions, surface slope, aspect, and other extraneous factors" (Lillesand 2004).

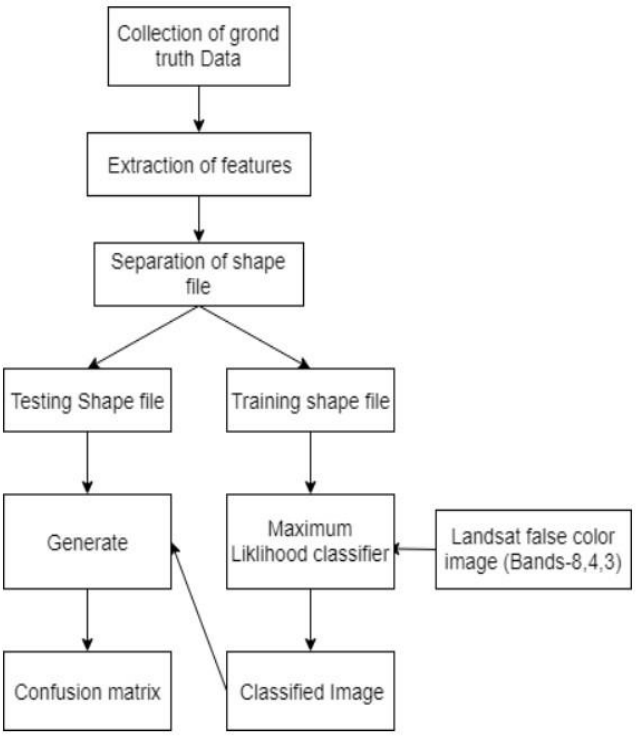

Figure 1: Generation of classified image

We aim to create a system where we can make prediction for the overall yield of the crop. The predictions would be for upcoming year. The major crop dependent for our paper is

sugarcane. The regression analysis takes GNDVI values and predicts the output between 0 to 1

To make our system more efficient we base our data on the ground truth points collected through a survey. The survey included relevant information like the crop name, age of the crop and previous crop

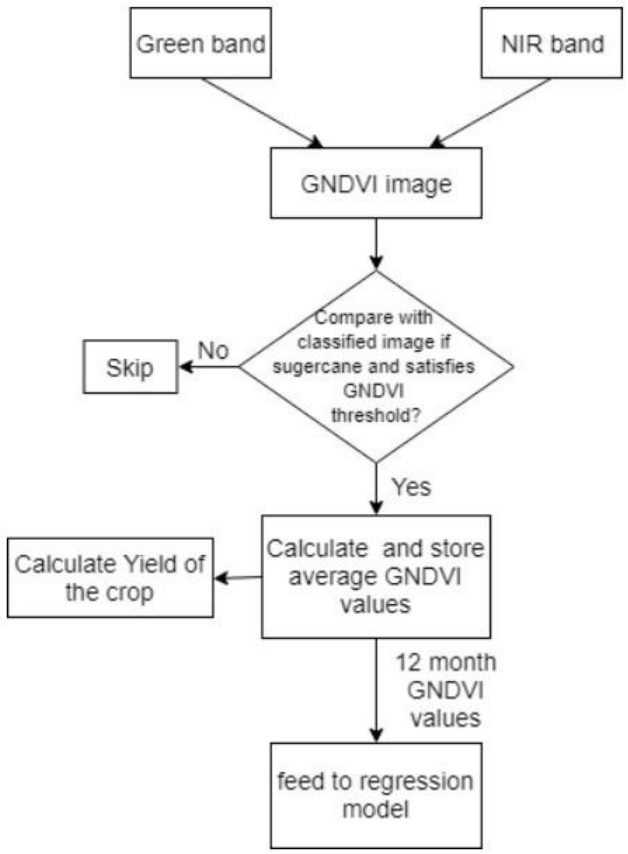

Figure 2: Calculation of average GNDVI values and yield for one crop cycle.

Study Area:

The study was carried out at Bagalkot cane growing region located in the situated-on branch of River Ghataprabha in Karnataka State of India. The area is located at 16.1817°N 75.6958°E, covering an area of 49.06 km2. The soil type in this area varied enormously due to climate, substance of parent material and topography. The mean annual rainfall of the area was recorded to 318 mm in 2016. The region received less rainfall than usual as a result of that farmers faced drought.

Implementation:

A. Generate classified image:

1. *collection of ground truth points:*

The survey was carried out for sugarcane fields in Sameer-wadi. The survey consists of capturing sugarcane field locations and crops surrounding the target fields. These points were used as the basis for the classification of the images. The collection of ground-truth data enables calibration of remote-sensing data, and aids in the interpretation and analysis of what is being sensed [2].

2. *Feature extraction:*

The ground truth tracks are in the form of polygons, these polygons are georeferenced and placed on the satellite image. After georeferencing these files, they are converted into shape files which contain labelled data and georeferencing information. These files are later used for classification process.

3. *separation of shapefiles:*

The maximum likelihood supervised learning algorithm in ArcGIS is used for classification. As we know that supervised learning required two types of input i.e. training and testing data. The shape files are divided into two types data in training sample manager, 70% data are given as training data and remaining 30% data as testing data to the algorithm. The training data is used to train the system which would in turn help to provide us the required results for the crop. The testing data helps to identify whether the implemented algorithm works properly.

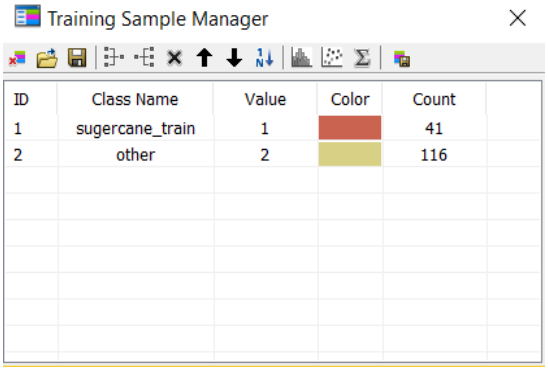

Figure 3: Separation of shapefile.

4. *Classification in ArcGIS:*

Landsat 8 imagery consists of 11 bands out of which three bands are used namely NIR (Near Infrared), Red and Green. The composite false color image is formed using these bands and provided as input for the classification process. Maximum Likelihood supervised learning algorithm is applied for classification in ArcGIS. The shape files generated in previous step are converted to signature files, it contains information about the target classes such as covariance and mean. The input to the algorithm consists of a composite raster image and a signature file. The basic formula that the algorithm follows is given as:

$$P(v|\hat{\omega}_i) = (2\pi)^{\frac{-N}{2}} |\hat{\Sigma}_i|^{\frac{-1}{2}} \exp\left\{\frac{-1}{2}(v-\hat{m}_i)^T \hat{\Sigma}_i^{-1}(v-\hat{m}_i)\right\}$$

$P(v|\hat{\omega}_i)$ is known from training data.
N dimensional space.
$\hat{m}_i$ and $\hat{\Sigma}_i$ is mean vector and covariance matrix of the data in class $\hat{\omega}_i$.

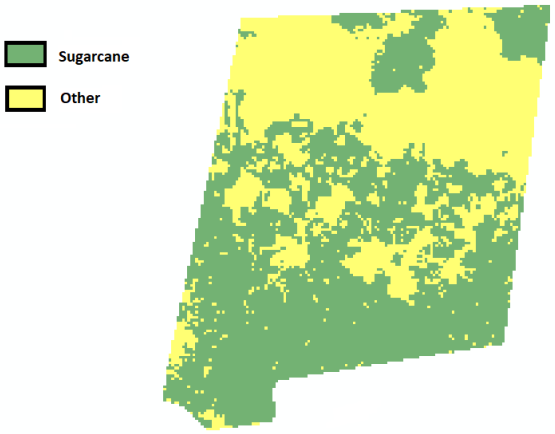

Figure 4: Classified image

5. *Confusion Matrix:*

This involves identifying a set of sample locations that are visited in the field. The ground truth point collected from the field is then compared to that which was mapped in the image for the same location. The comparison is done by generating confusion matrix (error matrix) [3]. The classified image and testing shape file are given input to the confusion matrix tool. In predictive analytics, a table of confusion (sometimes also called a confusion matrix), is a table with two rows and two columns that reports the number of false positives, false negatives, true positives, and true negatives. The figure-5 shows that how many time sugarcane class is identify as sugarcane class and how many other class identify as other class. Which finally results in the accuracy of the algorithm used in ArcGIS.

ClassValue	Sugercane	Other	Total	U_Accuracy	Kappa
Sugercane	216.00000000000	0.00000000000	216.00000000000	1.00000000000	0.00000000000
other	17.00000000000	267.00000000000	284.00000000000	0.94014084507	0.00000000000
Total	233.00000000000	267.00000000000	500.00000000000	0.00000000000	0.00000000000
P_Accuracy	0.92703862661	1.00000000000	0.00000000000	0.96600000000	0.00000000000
Kappa	0.00000000000	0.00000000000	0.00000000000	0.00000000000	0.93136526598

Kappa coefficient is the measure of agreement between two binary variables. If kappa coefficient equals to 0, there is no agreement between the classified image and the reference image. If kappa coefficient equals to 1, then the classified image and the ground truth image are totally identical [4]. So, the higher the kappa coefficient, the more accurate the classification is.

$$K = \frac{p_0 - p_e}{1 - p_e} = 1 - \frac{1 - p_0}{1 - p_e}$$

B. Regression analysis:

Regression analysis is widely used for prediction and forecasting, where its use has substantial overlap with the field of machine learning [5]. Green vegetation reflects more energy in the near- infrared band than in the visible range. Leaves reflects less in the near-infrared region when they are stressed, diseased or dead. Features like Clouds, water and snow show better reflection in the visible range then the near-infrared range, while the difference is almost zero for rock and bare soil. Regression analysis is also used to understand which among the independent variables are related to the dependent variable, and to explore the forms of these relationships GNDVI is a modified version of the NDVI to be more sensitive to the variation of chlorophyll content in the crop. It is useful for assessing the canopy variation in biomass and is an indicator of senescence in case of stress or late maturity

stage. This index can be used to analyze crops in mid to late growth stages.

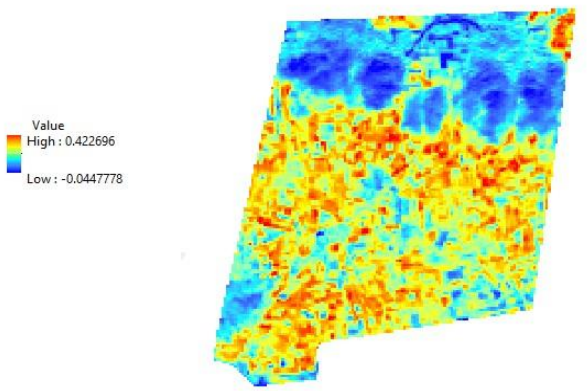

Figure 6: GNDVI image

NIR and green bands are used to form GNDVI image using the formula

$$GN\text{\textquestiondown}VI = \frac{(NIR - R)}{(NIR + R)} =$$

GNDVI values range from -1 to 1.

The figure-6 shows the GNDVI image which formed after applying GNDVI formula shown in above equation. After creating the GNDVI image it is compared with the classified image pixel by pixel, during this comparison the GNDVI values of area where sugarcane is identified are stored and later fed to the regression model to identify growth pattern, along with-it misclassification of rivers, barren land and urban structures such as roads and houses are removed by using a threshold on GNDVI value. It is known that values ranging from -1 to 0 does not denote vegetation but it is observed that barren lands and urban structures can have GNDVI values ranging between 0 to 0.25 hence a threshold to accept pixels or area having GNDVI values greater than 0.25 is applied. This classification provides yield in acres, it is calculated as

Yield (in acres) = number of pixels identified as sugarcane*900*0.000247105

The model derived GNDVI values were plotted over the calculated GNDVI values from Landsat images. The model was shifted vertically in each year to pass through the calculated GNDVI value acquired near or at the maximum period of sugarcane. The highest GNDVI value from the model was regressed against the final average crop yield measured in that year. From the model, the maximum GNDVI values for some months were 0.51, 0.58, 0.55 and 0.58 respectively.

The model represents maximum GNDVI in the month of September where the sugarcane yield is maximum for the year 2017

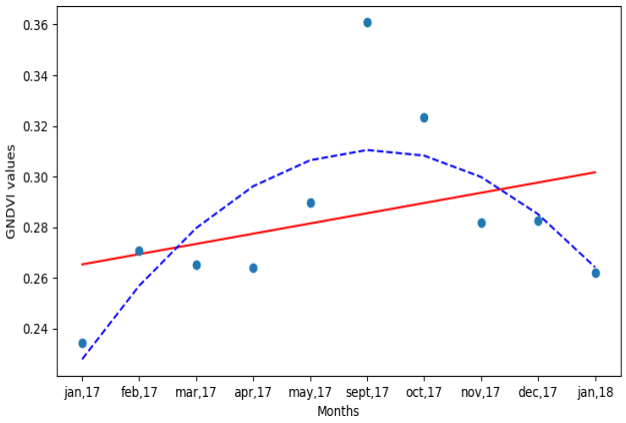

Figure 7: Regression graph

Conclusion:

Since the classified image produced by maximum likelihood classification has low accuracy, hence a threshold on GNDVI value is applied such that it rejects misclassification of barren lands, urban area, roads, rivers and lakes. The plant growth pattern of one year (2017-2018) is identified as shown in figure-7, as observed the plant growth increases and then after some amount of time it decreases again this denotes a complete cycle of sugarcane crop growth, the decreasing values show that the crop has reached its maturity and is ready to be harvested. The calculated yield based on one-year analysis is 1643(in acres) whereas the actual observed yield is 1409(in acres), it means there is an overestimation of 234(in acres). To minimize this error large amount of historical data regarding actual yield and satellite images are required. Due to lack of resources the analysis period in this paper is limited to one year (2017-2018) which is not enough data to precisely analyses the relation between yield and GNDVI values.

Future scope:

Provided that large amount of historical data is available regarding yield and satellite imagery, it is possible to identify crop growing patterns for better accuracy. The average GNDVI values can be plotted against observed yield to study the variation in yield due to changing weather conditions which affect the plant growth and in turn GNDVI values. Various other machine learning techniques can also be used to learn the growth pattern.

Limitations:

The supervised algorithm maximum likelihood does not provide accurate results; hence it affects the approximation values obtained while comparing classified/ observed values with actual values. Also lack of data/ resources becomes a hurdle as machine learning require huge amount of data for learning.

References:

1. Camelia Slave. 2014. Analysis of Agricultural Areas Using Satellite Images
2. David B. Lobell and Gregory P. Asner. 2003. Comparison of Earth Observing-1 ALI and Landsat ETM+ for Crop Identification and Yield Prediction in Mexico
3. Vricha Chavan, Shyamal Virnodkar. R N Awale A Review of Classification of Crops using Satellite Images and Applications in Estimation of Yield
4. N. GökhanKasapoglu,and Okan K. Ersoy. 2007. Border Vector Detection and Adaptation for Classification of Multispectral and Hyperspectral Remote Sensing Images
5. Andrew J. Robson. 2016. Prediction Using Landsat Time Series Imagery: A Case Study on Bundaberg Region