Surface Water Body Mapping In Himalayan Region

Palakurthy Dheeraj Srivathsav
Graduation in Data science and Analytics
dheerajsrivathsav@ou.edu
Date: 05-10-2018

Abstract

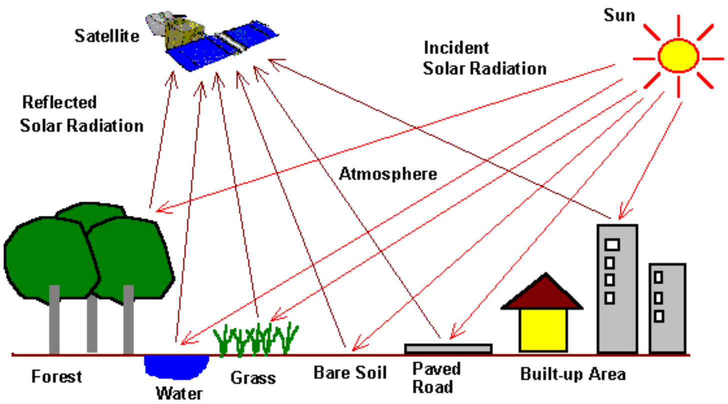
The planet is warming, from North Pole to South Pole. Since 1906, the global average surface temperature has increased between 1.1 and 1.6 degrees Fahrenheit (0.6 to 0.9 degrees Celsius)–even more in sensitive Polar Regions. The planet is already suffering from some impacts of global warming like Ice is melting worldwide, especially at the Earth's poles, many species have been impacted by rising temperatures, the sea level has been rising more quickly over the last century, and Precipitation (rain and snowfall) has increased across the globe, on average. Hurricanes and other storms are likely to become stronger. Floods and droughts will become more common. And there are lot of other consequences which the planet is going to face in the future because of global warming. Global warming appears to be rapidly accelerating temperature elevations in certain high-altitude zones, particularly those nearest the equator, which includes the Greater Himalayas. The melting of these Himalayan glaciers will have a more direct long-term effect on the hundreds of millions of people who live along rivers fed by their seasonal runoff. Hence in this project I find the opportunity to observe the effects of the Global warming on the Himalayas. For this I have selected two study areas named Wular Lake and the Dal Lake which are in very very close(less than 30 miles) from the Himalayan Mountain Ranges. The major water source for this lake will be the from the newly formed glaciers which happens due to melting of ice. So, by observing the surface area of these lakes over a period of time will indicate the rate at which the ice is melting, and the affect Global Warming has on Himalayas. For this project I have done two tasks. In the first task I have compared two different algorithms for the surface water body mapping one using MNDWI vegetative Index and the other using Machine Learning classification techniques like Random forest and SVM (Support vector machines) to predict the water body surface for two regions Wullar lake and the Dal lake. In the second task I have done the Time Series Analysis of Wullar lake from 2013-2017 using MNDWI Vegetative Index to observe the increase in the surface area of the lake. I have found that with the increase in the no.of training samples for both the classification techniques (Random Forest and SVM) gave us the better results because the model is gets trained well.For the time series analysis of Wullar lake the surface area of the is increasing from 2013 to 2017 except for 2016. This increase in the area is because of the melting of the ice. So, because of increase in global warming the ice is melting at the faster rate. Hence, this study projects how important it is for us to protect the environment by decreasing the global warming.

Introduction:

Optical remote sensing makes use of visible, near infrared and short-wave infrared sensors to form images of the earth's surface by detecting the solar radiation reflected from targets on the ground. Different materials reflect and absorb differently at different wavelengths. Thus, the targets can be differentiated by their spectral reflectance signatures in the remotely sensed images. Optical remote sensing systems are classified into the following types, depending on the number of spectral bands used in the imaging process.

We have interpreted bandwidths of different satellites sensors and analyzed how much reflectance is calculated by different sensors for various field of views. Later we have calculated the vegetative indices like NDVI (Normalized difference vegetation index), EVI (Enhanced vegetation index), LSWI (Land Surface Water Index) for every field of view with respect to each and every sensor.

After all this analysis we have concluded how different targets react to different wavelengths and obtain vegetative Index values for every satellite sensor. We have listed down the factors that has affected the plants in all these different conditions.



- Panchromatic imaging system: The sensor is a single channel detector sensitive to radiation within a broad wavelength range. If the wavelength range coincide with the visible range, then the resulting image resembles a "black-and-white" photograph taken from space. The physical quantity being measured is the apparent brightness of the targets. The spectral information or "color" of the targets is lost. Examples of panchromatic imaging systems are: IKONOS PAN, SPOT HRV-PAN
- Multispectral imaging system: The sensor is a multichannel detector with a few spectral bands. Each channel is sensitive to radiation within a narrow wavelength band. The resulting image is a multilayer image which contains both the brightness and spectral (color) information of the targets being observed. Examples of multispectral systems are: LANDSAT MSS, LANDSAT TM, SPOT HRS-XS, IKONOS MS

- Super spectral Imaging Systems: A super spectral imaging sensor has many more spectral channels (typically >10) than a multispectral sensor. The bands have narrower bandwidths, enabling the finer spectral characteristics of the targets to be captured by the sensor. Examples of super spectral systems are: MODIS, MERIS

- Hyperspectral Imaging Systems: A hyperspectral imaging system is also known as an "imaging spectrometer" it acquires images in about a hundred or more contiguous spectral bands. The precise spectral information contained in a hyperspectral image enables better characterization and identification of targets. Hyperspectral images have potential applications in such fields as precision agriculture (e.g. monitoring the types, health, moisture status and maturity of crops), coastal management (e.g. monitoring of phytoplankton’s, pollution, bathymetry changes). An example of a hyperspectral system is: Hyperion on EO1 satellite

Scientific questions:

Question-1:

The major question which I had during my project was that how to figure out the set a water pixels which remained same over the years and the water pixels which are newly formed in the recent years. So initially I didn’t understood how to differentiate between the water pixels and hence have considered all the water pixels for my analysis. That is I have calculated the area using all the water pixels and have comparted these areas over my study duration.

Question-2:

For the second method of analysis where I have to calculate the water body surface area Using vegetative Index. I have only used mNDWI as the vegetative index and have set the threshold of >0 to be water pixel and <0 to be a non-water pixel. I have set the threshold by observing lot of images and the ranges of water pixels in that image.

Objective of the study:

As I have mentioned earlier, Global warming appears to be rapidly accelerating temperature elevations in certain high-altitude zones, particularly those nearest the equator, which includes the Greater Himalayas. The melting of these Himalayan glaciers will have a more direct long-term effect on the hundreds of millions of people who live along rivers fed by their seasonal runoff. Scientists warn that by 2070 there could be a 43 percent decrease in land mass covered by these once seemingly eternal aspects of our planet’s architecture

Hence in this project I find the opportunity to observe the effects of the Global warming on the Himalayas. For this I have selected two study areas named Wular Lake and the Dal Lake which are in very very close(less than 30 miles) from the Himalayan Mountain Ranges. The major water source for this lake will be from the newly formed glaciers which happens due to melting of ice. So, by observing the surface area of these lakes over a period of time will indicate the rate at which the ice is melting and the affect Global Warming has on Himalayas.

2. Materials and Methods:

2.1 Study sites:

a) Wular Lake:

Wular Lake (also spelt **Wullar**) is the largest lake in India and one of the largest fresh water lakes in Asia. It is sited in Bandipora district in the Indian state of Jammu and Kashmir. The lake basin was formed as a result of tectonic activity and is fed by the Jhelum River. The lake's size varies seasonally from 12 to 100 square miles (30 to 260 square kilometers).

Fig 2.1.Location of Wular lake on India map

Fig 2.2 Melting of ice from Himalayas and formation of new glaciers

b) **Dal Lake:**

Dal is a lake in Srinagar (Dal Lake is a misnomer as Dal in Kashmiri means lake), the summer capital of Jammu and Kashmir. The urban lake, which is the second largest in the state, is integral to tourism and recreation in Kashmir and is named the "Jewel in the crown of Kashmir" or "Srinagar's Jewel". The lake is also an important source for commercial operations in fishing and water plant harvesting. This lake gets frozen up in the winter.

Fig 2.3. Dal lake

Fig. 2.4 Scenic view of Dal lake

Fig. 2.5 Frozen Dal Lake during winter

2.6 children playing on Frozen Dal Lake

2.2 Landsat data:

The study area on which I am working does not have much of good satellite images as it was very close to the mountain ranges. Hence I have used three satellite sensors LANDSAT 5, LANDSAT 7 and LANDSAT 8 for the image collection and used the one best image during a specific time interval. For the Time series analysis of 5 years I have used 5 best images during each time Interval.

Fig 2.7 Landsat 5 Image

2.3 Statistical Analysis and Mapping algorithms or models:

For the task 1 I have used Machine Learning classification techniques like Random forest and SVM (Support vector machines).

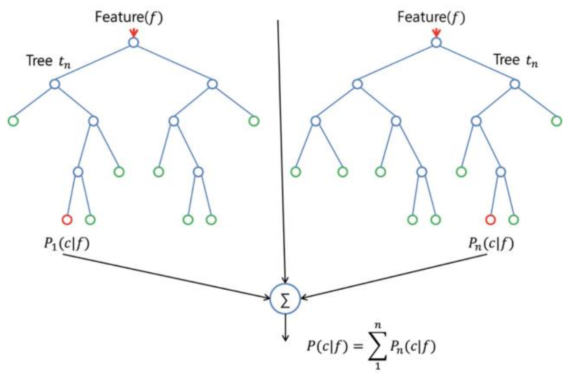
Random Forest:

Random Forest is a flexible, easy to use machine learning algorithm that produces, even without hyper-parameter tuning, a great result most of the time. It is also one of the most used algorithms, because its simplicity and the fact that it can be used for both classification and regression tasks. In this post, you are going to learn, how the random forest algorithm works and several other important things about it.

How it works:

Random Forest is a supervised learning algorithm. Like you can already see from it's name, it creates a forest and makes it somehow random. The „forest" it builds, is an ensemble of Decision Trees, most of the time trained with the "bagging" method. The general idea of the bagging method is that a combination of learning models increases the overall result.

One big advantage of random forest is, that it can be used for both classification and regression problems, which form the majority of current machine learning systems. I will talk about random forest in classification, since classification is sometimes considered the building block of machine learning. Below you can see how a random forest would look like with two trees:



With a few exceptions a random-forest classifier has all the hyper parameters of a decision-tree classifier and also all the hyper parameters of a bagging classifier, to control the ensemble itself. Instead of building a bagging-classifier and passing it into a decision-tree-classifier, you can just use the random-forest classifier class, which is more convenient and optimized for decision trees. Note that there is also a random-forest regressor for regression tasks.

The random-forest algorithm brings extra randomness into the model, when it is growing the trees. Instead of searching for the best feature while splitting a node, it searches for the best feature among a random subset of features. This process creates a wide diversity, which generally results in a better model.

Support Vector Machines (SVM):

Support vector machines (SVMs) are a set of supervised learning methods used for classification, regression and outlier's detection.

The advantages of support vector machines are:

- Effective in high dimensional spaces.
- Still effective in cases where number of dimensions is greater than the number of samples.
- Uses a subset of training points in the decision function (called support vectors), so it is also memory efficient.

- Versatile: different Kernel functions can be specified for the decision function. Common kernels are provided, but it is also possible to specify custom kernels.

3. Results:

3.1 Image selection

There were about 30 images available from Landsat 8 for Wullar Lake in 2017. Some of the images were poor quality due to the presence of clouds. Based on image observation, six best quality images are presented in Fig. 3.1 to Fig 3.6. For the further analysis I used the cloud free best quality image shown in Fig. 3.6.

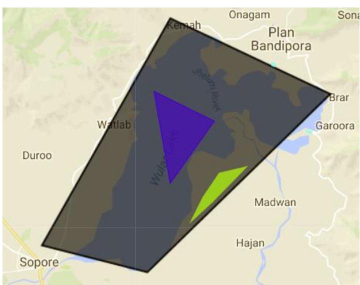


Fig 3.1 Original Image

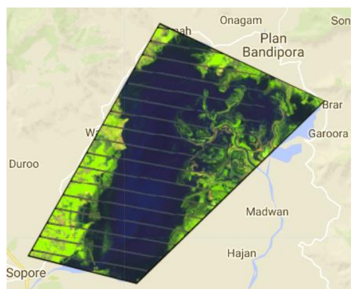


Fig 3.2 Landsat 5 Image

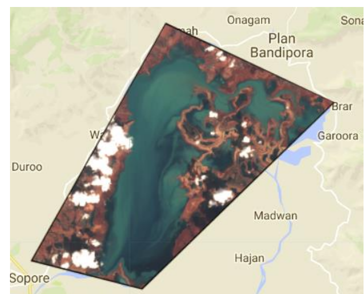


Fig 3.3 Landsat 7 Image

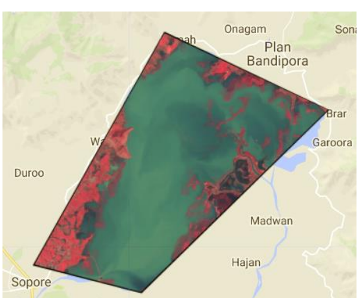


Fig 3.4 Landsat 8 Image

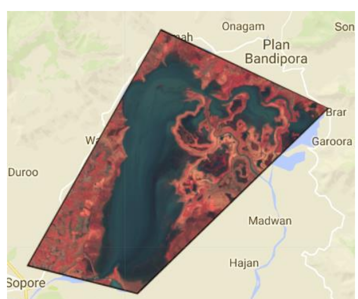


Fig 3.5 Landsat 8 Image

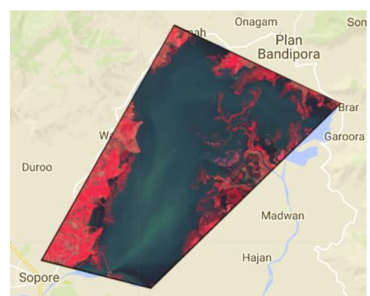


Fig 3.6 Landsat 8 Image(Best Image)

3.2 Random Forest classification

Fig. 3.7 shows the training samples for different landcover types (water, grassland and urban) in the study area.The land sat-8 image for the selected region is shown in fig 3.8. Fig. 3.9 shows the image in this which I have trained the model using the Random forest classifier and it was not able to detect the grassland portion as the training sample is very very small.The water content after classification is represented by blue color and the urban are is representedby red color. Fig. 3.10 results after classification using Random Forest and converting it to a Binary Image

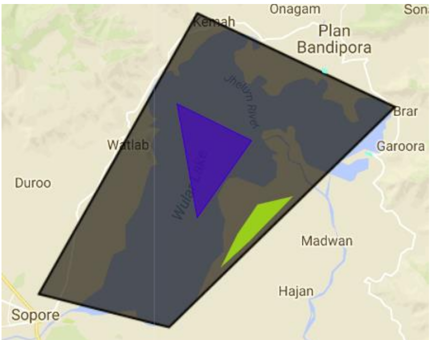


Fig 3.7 Original Image

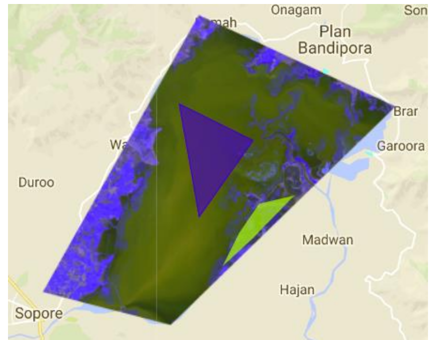


Fig 3.8 Landsat-8 Image

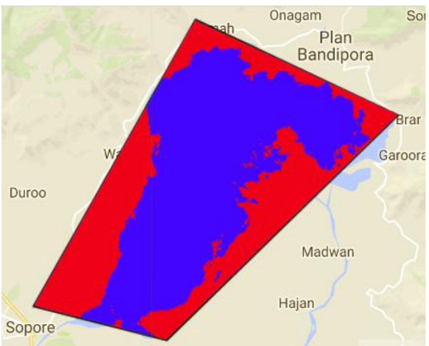


Fig 3.9 Image After Classification

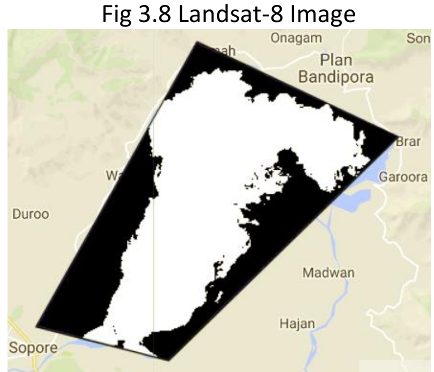


Fig 3.10 Binary Image

3.3 Support Vector Machine(SVM) classification

Fig. 3.11 shows the training samples for different landcover types (water, grassland and urban) in the study area.The land sat-8 image for the selected region is shown in fig 3.12. Fig. 3.13 shows the image in this which I have trained the model using the SVM classifier and even this classifier was not able to detect the grassland portion as the training sample is very very small.The water content after classification is represented by blue color and the urban are is representedby red color. Fig. 3.14 results after classification using Random Forest and converting it to a Binary Image

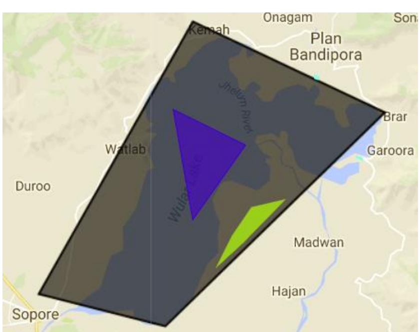


Fig 3.11 Original Image

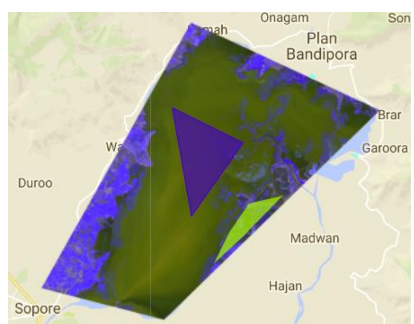


Fig 3.12 Landsat-8 Image

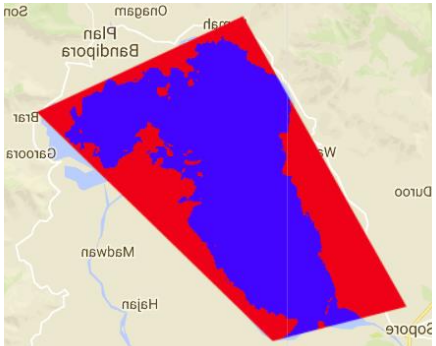


Fig 3.13 Image After Classification

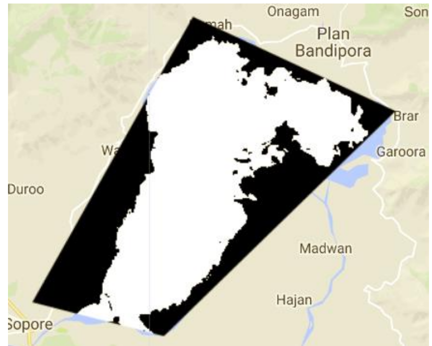


Fig 3.14 Binary Image

3.4 Random forest VS SVM for basic Training Samples

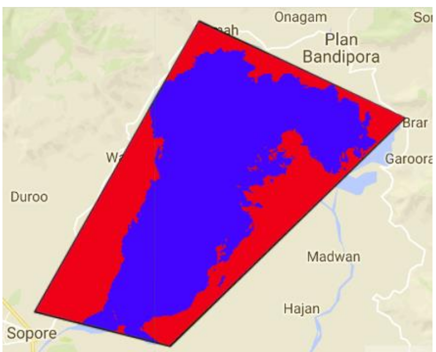


Fig 3.14 Classification using Random Forest

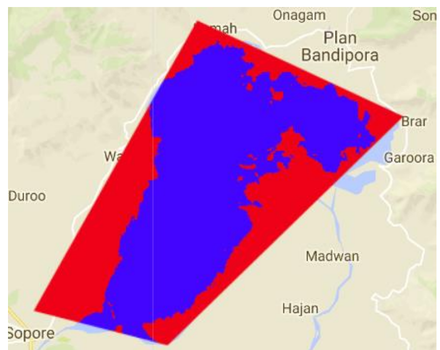


Fig 3.15 Classification using SVM

Training Samples:
Water: 1 polygon
Grassland: 1 Polygon
Forest: 2 Polygons
Area of the surface water: 102011.
Water: 3 polygon
Grassland: 3 Polygon
Forest: 2 Polygons
Area of the surface water: 102015.4

Training Samples:
Water: 1 polygon
Grassland: 1 Polygon
Forest: 2 Polygons
Area of the surface water: 103148.79
Water: 3 polygon
Grassland: 3 Polygon
Forest: 2 Polygons
Area of the surface water: 103291.91

The area estimated from both methods are compared in Fig.3.14 and Fig.3.15 With the same number of samples, the SVM showed larger area of the water than the RF method. However, results showed that in both methods increasing the number of training samples increased the water surface area.

3.5 surface water body mapping using MNDWI vegetative Index

Here in this method I have calculated the MNDWI vegetative Index of the study area and set the threshold value by Observing different pixels of water body and the non-water body and considered values with MNDWI values greater than zero as water body and less than zero as non-water body.

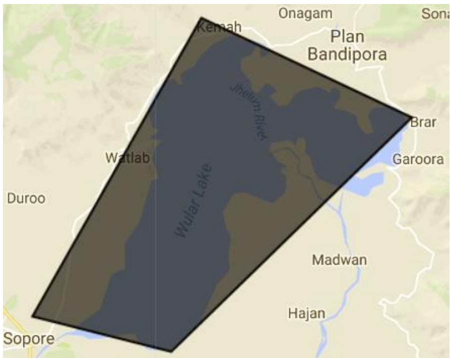


Fig 3.16 Original Image

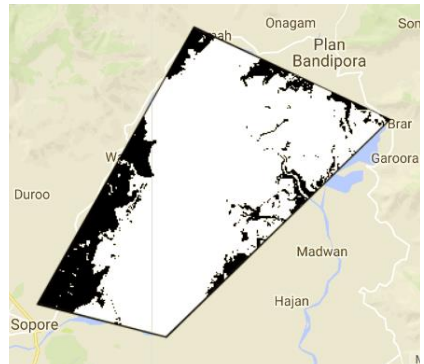


Fig 3.14 Binary Image after setting Threshold values

3.6 Comparison of the Algorithms:

Mapping the Water surface by Predicting the image using the Training Samples

Classifiers: Random Forest, SVM

Mapping the Water surface by calcuating the MNDWI values by setting the threshold values

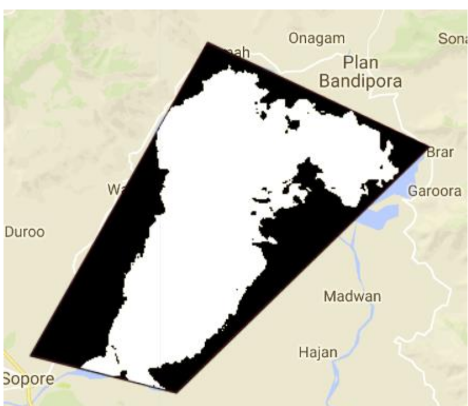


Fig 3.17 Binary Image of classification

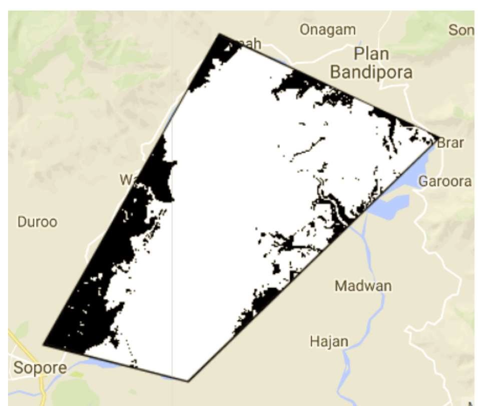


Fig 3.18 Binary Image using MDWI method

Area of the surface water: 103291.91

Area of the surface water: 164338.50

So, by comparing both the images I have found that the surface area of the water is more using vegetatative index methos compared with the classification technique.To verify my result I have applied the same methods for the DAL Lake.

3.7 Evaluating the same Task for different study area: Dal Lake

Repeated the 3.1 step to select the best Landsat images similar to the before study area (Wullar Lake)

Fig. 3.18 shows the training samples for different landcover types (water, grassland and urban) in the study area.The land sat-8 image for the selected region is shown in fig 3.19. Fig. 3.20 shows the image in this which I have trained the model using the Random forest classifier and it was able to detect all the vegetation land covers.The water content after classification is represented by blue color and the urban are is representedby red color. Fig. 3.21 results after classification using Random Forest and converting it to a Binary Image

Fig 3.18 Original Image

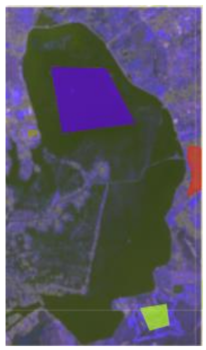


Fig 3.19 Landsat Image

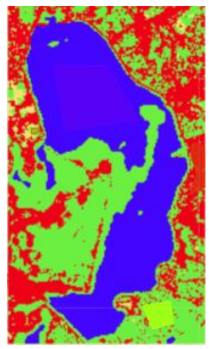


Fig 3.20 Classified Image


Fig 3.21 Binary Image

Step-3: Random forest VS SVM for basic Training Samples

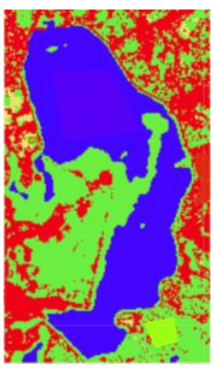


Fig 3.22 classification using Random forest with Minimum training samples

Training Samples:
Water: 1 polygon
Urban: 1 Polygon
Grassland: 1 Polygon
Forest: 1 Polygons
Area of the surface water: 12799.38

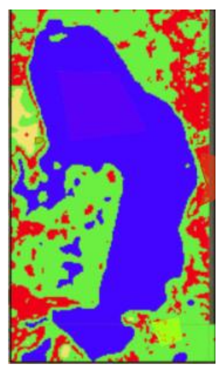
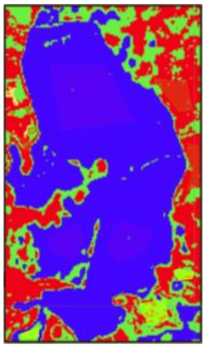
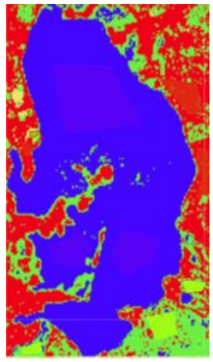


Fig 3.23 classification using SVM with Minimum training samples

Training Samples:
Water: 1 polygon
Urban : 1 Poygon
Grassland: 1 Polygon
Forest: 2 Polygons
Area of the surface water: 102011.4



Training Samples:
Water: 3 polygon
Urban: 3Polygon
Grassland: 3 Polygon
Forest: 2 Polygons
Area of the surface water: 20380.00



Training Samples:
Water: 3 polygon
Urban: 3Polygon
Grassland: 3 Polygon
Forest: 2 Polygons
Area of the surface water: 16272.64

Result of Increase in the Training Samples:

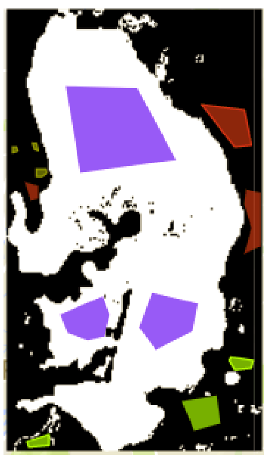


Fig 3.27 Binary Image with more Training samples

Training Samples:
Water: 1 polygon
Urban: 1 Polygon
Grassland: 1 Polygon
Forest: 1 Polygons
Area of the surface water: **20380.00**

Fig 3.26 Binary Image with basic Training samples

Training Samples:
Water: 3 polygon
Urban: 3 Polygon
Grassland: 3 Polygon
Forest: 2 Polygons
Area of the surface water: 12799.38

The area estimated from both methods are compared in Fig.3.27 and Fig.3.26 With the same number of samples, the Random Forest showed larger area of the water than the SVM method. However, results showed that in both methods increasing the number of training samples increased the water surface area.

3.8 surface water body mapping using MNDWI vegetative Index

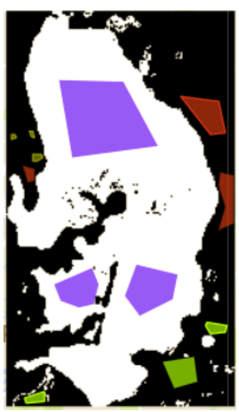
Here in this method I have calculated the MNDWI vegetative Index of the study area and set the threshold value by Observing different pixels of water body and the non-water body and considered values with MNDWI values greaterthan zero as water body and less than zero as non-water body.

Fig 3.28 Original Image

Fig 3.28 Binary Image after setting the threshold values

Comparison of the Algorithms:

Mapping the Water surface by Predicting the image using the Training Samples Classifiers: Random Forest, SVM



Area of the surface water: **20380.00**

Mapping the Water surface by calculating the MNDWI values by setting the Threshold values

Area of the surface water: 25048.23

So, by comparing both the images I have found that the surface area of the water is more using vegetatative index methos compared with the classification technique.To verify my result I have applied the same methods for the DAL Lake.

3.9 Based on comaparisons among three methods for two lakes, the MNDWI Vegetative Index methods estimated the surface water area better.Therefore, Time Series Analysis of Wullar lake from 2013-2017 using MNDWI Vegetative Index are presented in Fig. 3.29 to 3.33.

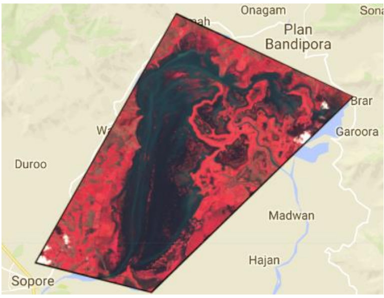


Fig 3.29 Best Image in 2013

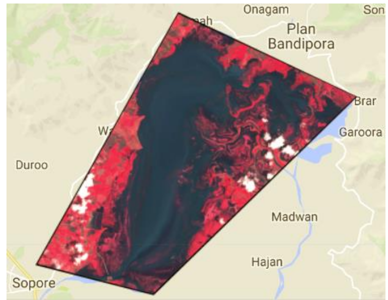


Fig 3.30 Best Image in 2014

Fig 3.31 Best Image in 2015

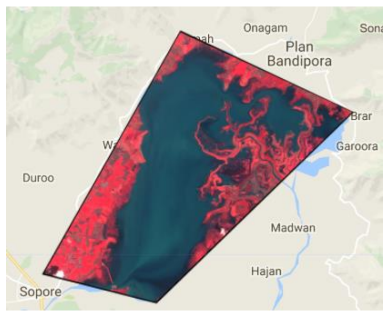


Fig 3.32 Best Image in 2016

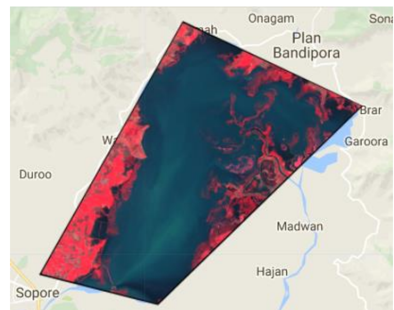
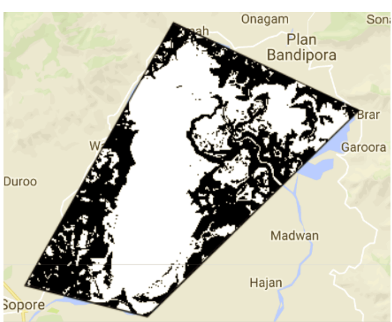


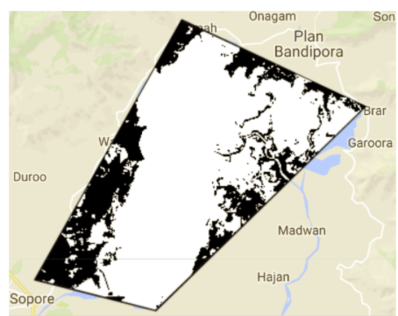
Fig 3.33 Best Image in 2017

3.10 Water Surface areas of each best Image:

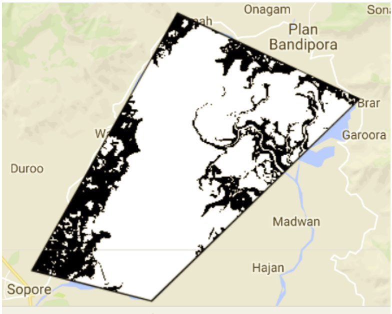
The water surface area in of the image is calucated and is presented from Fig of year 2013 to 2017

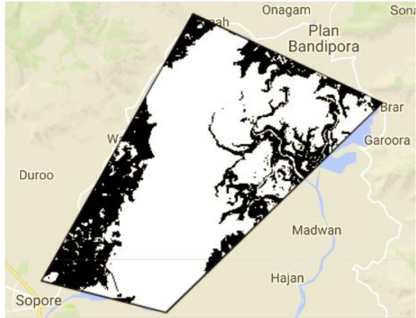


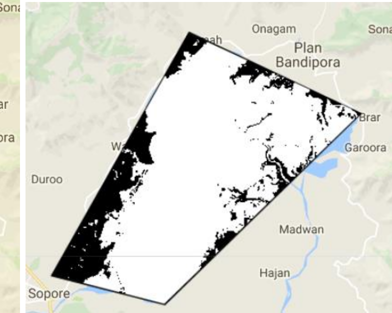
2013: Area of the surface water: **122173.03**



2014: Area of the surface water: 1515107.68

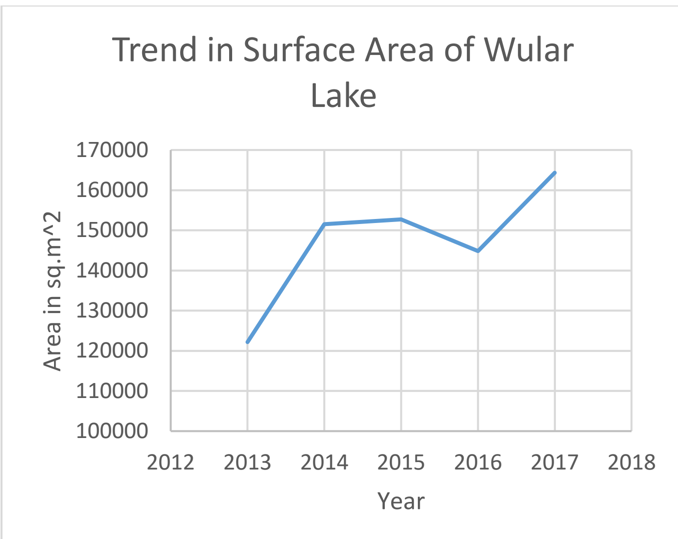






2015: Area of the surface water: **152759.00** **2016:** Area: **152759.00** **2017:** Area: **164855.30**

Chart: Trend in the surface area of Wular lake:



We can see that there is an increase in area of the surface water level throughout the 5 years except for year 2016 where there is a decrease in the trend. In 2013, the area estimated was about 122 thousands sq. km and reached to the total area of about 164 thousands sq. km in 2017. However, the area was about 144 tousands sq. km in 2016 which decreased from 152 thousands sq km in 2015.The reason for this unexpected decrease in 2016 could beless availability of proper satellite Image in this year. We can see from the above plots that the best satellite image of 2016 has a lot of variation with respect to the original image.

4.Discussion:

In this project I have implemented both the algorithms in a basic way.But as I have mentioned that I had a few questions(mentioned in the Scientic question section) while I was doing this project. But most of my questions were resolved after reading this paper "Continued decrease of open surface water body area in Oklahoma during 1984–2015 (Zhenhua Zou, Jinwei Dong, Michael A. Menarguez , Xiangming Xiao , Yuanwei Qin ,Russell B. Doughty, Katherine V. Hooker, K. David Hambright) ".This paper was very clear and has shown effective methods to implement surface water body mapping.In this paper they explored four indicators of surface water body extents based on the annual water body frequency: 1) the maximum water body extent in a given year, 2) the persistent year-long water body extent, 3) seasonal changes in water body area, which is the difference between the maximum and year-long water body extents, and 4) the annual average water body extent.

For each pixel, they counted the number of observations within a year it was identified as open surface water body, and then divided it by the total number of good observations in that year and termed the resultant ratio as water body frequency. For each year, they generated annual maps of maximum water bodies (water body frequency ≥ 0.25%), year-long water bodies (water body frequency ≥ 75%) since they have the water most of the year, and seasonal water bodies (25% ≤ water body frequency <= 75%) respectively, and then calculate the areas of maximum, year-long and seasonal water bodies correspondingly.

They have even mentioned the way to solve my question 2.They have used an efficient way to calculate the surface area of water. Despite the advantage of mNDWI over NDWI in the remote sensing of water, the mNDWI approach still has commission error in those mixed pixels of water and other land cover types. In particular, vegetation over a wet surface is one of the major causes for commission error in open surface water body mapping. In this study, they have combined mNDWI and vegetation indices (NDVI and EVI) to reduce the effects of vegetation on water body mapping algorithm. Specifically, they detected only pixels with water signal that was stronger than the vegetation signal as actual water pixels (mNDWI N NDVI or mNDWI N EVI). In order to further remove the noise caused by vegetation, EVI was applied to exclude the wetland pixels with vegetation (EVI b 0.1). Therefore, only those pixels that met the criteria ((mNDWI N NDVI or mNDWI N EVI) and (EVI b 0.1)) were classified as open surface water body pixels. The remaining pixels were classified as non-water pixels (SOM 5).

4.1Limitations of the study:

I haven't implemented these two methods in my project but have found a way to rectify my mistakes. If I would have got more time available, I would definitely have done better.

4.2 TASK-1: Areas of Improving the Model

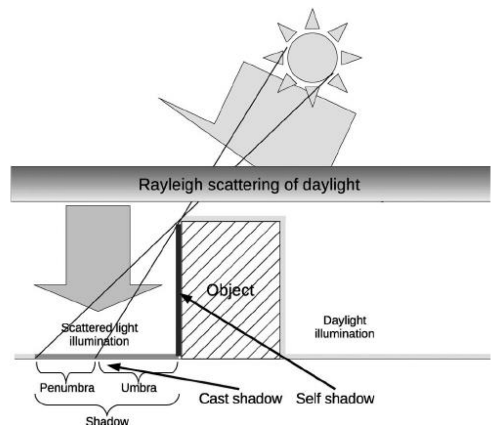
- The accuracy of the classifier increases with increase in the training samples. The more the training samples are the better the model gets trained and we there is a high chance of getting better predictions.
- The results of the classifier model can be improved in areas where there are good number of satellite images and also being most similar to the original Image.
- The major area of improving the results will happen when we apply Hyper parameter tuning and cross validation techniques to our classifier i.e by increasing the no.of iterations or tuning a particular parameter based on the technique we are using.

- There are several other classification techniques like Naive Bayes, Logistic Regression, Decision trees, Adaboost, Black Boost, and Artificial Neural Networks etc. By applying these methods we can verify whether there is an improvement in the results or not.

4.2 TASK-2: Effects on satellite Images in the Mountain Terrain

There are several effects on the satellite Images in the Mountain Terrains. Some of the very important ones are**:**

a. Cast Shadows and Self Shadows:



b. Variation of Ground slope:

The variation of ground slope must be accounted for in both the solar and sky irradiance components

5. Conclusion:

TASK-1: Comparison of Algorithms

- I have observed the results for this task for two regions Wular Lake and the Dal Lake. Both of the results show that the Water body mapping done using the MNDWI vegetative Index is better than the water body mapping done using the Machine Learning Techniques to predict the water body based on the training samples.

Area of the surface water Using classifier: **20380.00**	Area of the surface water using MNDWI Vegetative Index: **25048.23**

TASK-2: Time Series Analysis of Wular Lake

- We can see that there is an increase in area of the surface water level throughout the 5 years except for year 2016 where there is a decrease in the trend. The reason for this unexpected decrease in 2016 is because there is no proper satellite Image in this year. We can see from the above plots that the best satellite image of 2016 has a lot of variation w.r.t the original image.

The major sources for increase in water level are either due to the melting of the Himalaya mountain rages due to global warming or due to rainfall. With melting of the Himalayas new glaciers are formed and they are get merged into these lakes.

6. Acknowledgments:

My whole-hearted thanks to my team mates who have supported me during the data collection and data analysis.

I would like to thank Professor Dr. Xiangming Xiao for his expert advice and encouragement throughout this difficult project, as well as all the TA's who have supported in all the stages. I need to specially thank Dr.Jie Wang and Dr. Rajan for their valuable advice on how to approach this project.

7. References:

1. Continued decrease of open surface water body area in Oklahoma during 1984–2015 (Zhenhua Zou, Jinwei Dong, Michael A. Menarguez , Xiangming Xiao , Yuanwei Qin ,Russell B. Doughty, Katherine V. Hooker, K. David Hambright) ".

2. Aherne, J., Larssen, T., Cosby, B.J., Dillon, P.J., 2006. Climate variability and forecasting surface water recovery from acidification: modelling drought-induced sulphate release from wetlands. Sci. Total Environ. 365 (1–3):186–199. http://dx.doi.org/10.1016/j.scitotenv.2006.02.041.

3. Gorelick, N., Hancher, M., Dixon, M., Llyushcheno, S., Thau, D., Moore, R., 2017, GoogleEarth Engine:Planetary-Scale geospatial analysis for everyone, Remote Sensing of Environment,202:18-27