# Development of an open source tool to remove cloud from hyperspectral satellite imagery using machine learning approach.

**Rohith Gilla[1], Salghuna NN[2], Rama Chandra Prasad[3]**

[1]Mahindra ÉcoleCentrale, Bahadurpally, Hyderabad, Telangana, India - 500001
[2]Lab for Spatial Informatics, International Institute of Information Technology, Gachibowli, Hyderabad, Telangana, India – 500032

*[1]gillarohith1@gmail.com, [2]salghuna@gmail.com, [3]rcprasad@iiit.ac.in*
***Keywords**: CHRIS; Sentinel, Hyperspectral, Reflectance, GUI,*

---

**Abstract:**
The advent of hyperspectral remote sensing facilitated the mapping and classification of various features of the earth at a very detailed level. This technology has been significantly adopted in different domains like forestry, agriculture, geology, and hydrology in distinguishing components or objects at a micro level. However, there exist challenges in using this technology and one among them is getting a cloud-free satellite data. Presence of clouds in satellite imagery makes the data unsuitable for mapping purposes. To combat this scenario, researchers have come up with different algorithms to remove clouds from satellite imagery for efficient utilization of data in generating a thematic map. In view of this context, in the present study an attempt has been made to remove clouds from hyperspectral data (CHRIS) using a machine learning approach and also to create a GUI for other researchers working in similar domain. To achieve this objective, the cloud region within CHRIS data was identified and values for those areas were replaced using multi-spectral Sentinel data acquired during the same time. As a first step, CHRIS hyperspectral images were imported using tifffile module, and then converted them to a numpy array where the size of the array was length of the image, width of the image, number of layers of the image. The study used JPEG 2000 format to load sentinel data files, Pillow module to load image and then converted them to numpy array. As the coordinates per pixel of multispectral (Sentinel) is different from per pixel of hyper spectral, CHRIS data was georeferenced with reference to Sentinel data to match the geographical coordinates of both the images using linear equations as

$$a1 = \min(latitude\ of\ hyperspectral) \quad y1 = \min(longitude\ of\ multispectral)$$
$$b1 = \min(longitude\ of\ hyperspectral) \quad x2 = \max(latitude\ of\ multispectral)$$
$$a2 = \max(latitude\ of\ hyperspectral) \quad y2 = \max(long\square tude\ of\ multispectral)$$
$$b2 = \max(longitude\ of\ hyperspectral)$$

Further to find the starting pixel of hyperspectral in multispectral image following equation was used

$$xpixel_{start} = \frac{(a1 - x1)}{(x2 - x1)/len(x)}$$

And to match per pixel latitude and longitude of hyperspectral with that of multispectral datal, following linear equations were used to get a relation.

$$step_x = (x2 - x1)/(a2 - a1)$$
$$step_y = (y2 - y1)/(b2 - b1)$$

The Sentinel data consisted of 13 bands and hyperspectral CHRIS data consists of 40 bands. As an initial step, 10 bands out of 13, of Sentinel data (retaining 20 m resolution) were converted to 40 bands and then later the values for the missing bands was generated using linear regression. To increase efficiency of algorithm the study used a moving window model for calculation of average

reflectance of CHRIS image i.e, each pixel reflectance value is the sum of nine pixels within the window, so given a pixel $x1, y1$ the average reflectance value will be

$$Average\ Reflectance = \frac{\begin{matrix} Ref(x1-1,y1-1)+Ref(x1-1,y1)+Ref(x1-1,y1+1)+ \\ Ref(x1,y1-1)+Ref(x1,y1)+Ref(x1,y1+1)+Ref(x1+1,y1+1)+ \\ Ref(x1+1,y1)+Ref(x1+1,y1-1) \end{matrix}}{9}$$

Finally, a machine learning model was used to split Sentinel data and create multi virtual layers with respect to CHRIS data and then estimate the pixel value from Sentinel data and replacing the values of cloud pixels of CHRIS data of similar location. Towards implementing this, the study used one of the basic methods of machine learning, i.e. Linear Regression, where reflectance values are given as target variable and wave lengths input variable and make a model by training on multispectral data and predicting for hyperspectral wavelengths. The study used scipy, scikit learn to build these models. After the creation of arrays by prediction, the array was converted into a layer and saved as Image, resulting to an output of 35+ images and one image containing all layers. To ease this operation for future researchers, GUI is created using python tkinter module. Within GUI there will be dialog boxes to choose input images, placeholders to give co-ordinates, a dialog box to choose destination, and all the output files will be saved in output folder which is created inside the destination folder.

## 1. Author/s Biography

I am Rohith Gilla from India, currently pursuing under graduation 3rd year in Computer Science Major at Mahindra Ecole Centrale, Hyderabad. I have completed my schooling in Vivekananda Residential School with a CGPA of 10 and did my high schooling in PAGE Institute with 98%. I am a passionate coder and re-searcher. I am an Open Source enthusiast and have done a good number of contributions in it and coming to my tech stack my major work is on Python and other stack includes Django, Django REST framerwork, REST, SOAP, flutter, Kotlin, C/C++, Java, Vue-js, bash,LaTeX, git and also a full stack web developer. I am an Open Source community manager of Coderplex in Hyderabad. Currently, I work with a startup as a software developer and also did internships at International Institute of Information Technology, VerZeo. I have the ability and enthusiasm to readily master new technologies. I've worked on projects like News analyser twitter Bot, College ShoutOut Box, Web scrappers, Spiders, CLI applications, PyConIndia Official twitter bot and many more, you can find them here https://github.com/Rohithgilla12/

Apart from this I volunteer in a NGO called Make a Difference which tries to ensure equitable outcomes for children in shelter homes.

N. N. Salghuna is a research scholar at Lab for Spatial Informatics, IIIT-Hyderabad, India. She has pursued her Master's in Birla Institute of Technology, Ranchi. Her research interest includes solving of social issues (land record management, drought management, storm mitigation) using geospatial

tools. Currently she works on forest species identification and assessing their health conditions using field-collected biochemical parameters and hyperspectral data, towards her Ph.D.

R. C. P. Pillutla works as assistant professor at Lab for Spatial Informatics, IIIT-Hyderabad. He has completed his Ph.D. in the Environmental Sciences under joint affiliation of National Remote Sensing Centre (NRSC-ISRO), and Department of Environmental Sciences, Osmania University, Hyderabad. His research interest includes management of natural resources (primarily forest) using geospatial tools and promoting awareness towards the utility of open source geospatial software.