Design of an algorithm to obtain and process hyperspectral images

1. Introduction
2. Theorical framework
    2.1 Light and wavelengths
3. Obtaining samples
4. Libraries and tools
5. Processing images
    5.1 Obtaining and analyzing samples
    5.2 Creating an AI model
6. Regards
7. Conclusions

1. Introduction

EMIT 4 Life is a web portal designed to display, in a clear and intuitive manner, all the information derived from the hyperspectral analysis of satellite images, in order to provide key information for the analysis and decision-making regarding biodiversity conservation.

The initial study area is the floral region of the Cape, South Africa, and the data will consist of images taken by the EMIT sensor on the International Space Station (as will be seen later). The objective of this article is to complement the project demo with a conceptual implementation of the algorithm that would be responsible for providing the information that would be displayed graphically on the Emit 4 Life portal.

2. Theorical framework

Spectral analysis of images is a very broad and complex branch of physics. In order to illustrate the basic functioning of the algorithm, a theoretical framework has been prepared regarding the formulas and analyses that will be employed later in the project.

2.1 Light and wavelengths

The white sunlight that reaches the Earth's surface is composed of a multitude of waves oscillating at different frequencies. Among all these waves, the ones of interest to us are the colors of the visible spectrum (from red to violet) and the infrared rays.

This is because when light reaches the surface and bounces back, some of these wavelengths are absorbed by different materials. Without delving into the behavior that causes this phenomenon, we can use the information from the wavelengths that bounce back into the atmosphere to deduce what material the light has collided with.

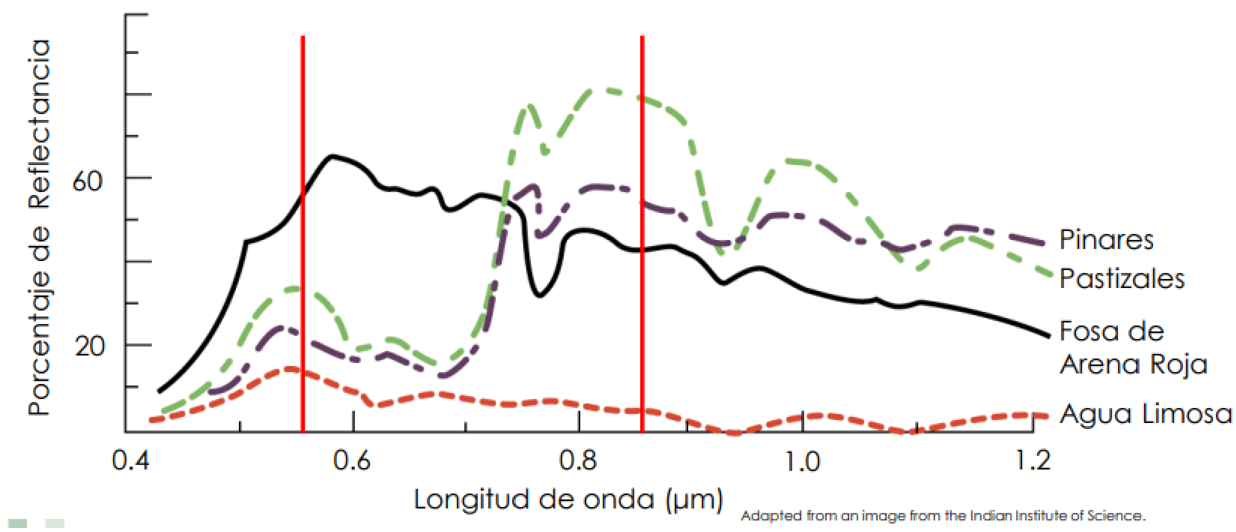

Figure 1. Wavelength absorption for different materials

In this way, by identifying reflectance patterns in different areas of an image, we can identify the materials present in it. This includes everything from mineral deposits to plant life and suspended particles in both air and water. This reflectance pattern is known as the spectroscopic signature of a material.

However, this reflectance can be affected by the absorption of the atmosphere, which also absorbs certain wavelengths. Nevertheless, this has not been taken into account because, as will be explained in point 3, the samples used have already been corrected to avoid this issue.

3. Obtaining samples

The samples have been obtained from the EarthData Search portal of NASA. They belong to the newly installed EMIT sensor on the International Space Station and correspond to an area known as the Cape Floral Region in South Africa.

The data is downloaded using the API of the Common Metadata Repository of the Earth Observing System. To do this, it is necessary to create an account with this organization and use the data request functions.

In the case of EMIT data, as mentioned earlier, these are not raw data; they have already been preprocessed by removing cloud layers and compensating for spectrum deformation caused by the atmosphere.

```
curl -v -i https://cmr.earthdata.nasa.gov/stac/LPCLOUD?page=2
```

4. Libraries and tools

For the implementation of the algorithm that will be explained later, we have used the following libraries and tools:

1. QGIS: It is a user-friendly, free, and open-source professional Geographic Information System (GIS) that enables the creation, visualization, analysis, editing, and publication of geospatial information.
2. GeoJSON: It is an open standard format designed to represent simple geographic elements along with their non-spatial attributes.
3. AVHYAS: This is a Python-based plugin for QGIS designed for the processing and analysis of hyperspectral images.

5. Processing images

Next, we proceed to explain the structure and specifications of the algorithm responsible for obtaining and processing the data that will be displayed on the web. The goal of this algorithm is not only to acquire the information, but also to transform it into a lighter and more manageable format for web browsers. This algorithm consists of two parts: data retrieval and data processing.

5.1 Obtaining and analyzing samples

As briefly explained in section 3, we have used the CMR API with the POST method to be able to use shapefiles in order to delineate the area of our study. Therefore, if we want to study, for example, the Amazon, we would only need to provide the algorithm with a shapefile composed of points or polygons in the 'shape' parameter. Through the API, we would obtain a list of links which we will store in a list, a list that we will in turn use to keep track of which granules are still pending processing. The algorithm will iterate through the list one by one, taking each link, downloading the corresponding granule, and processing it with our classification model. Additionally, there are two files, 'pending' and 'done'. The 'pending' file is a pickle file that contains the list of granules pending for download and processing, while the 'done' file is simply a text file where the successfully processed granules are recorded.

This script results in a number of processed and classified TIFF images corresponding to the number of granules.

5.2 Creating an AI model

We have utilized QGIS, which can be automated with Python, to train an artificial intelligence model, specifically a supervised classification model. The initial step involves converting the NetCDF4 image to ENVI format, as NetCDF4 cannot be used to train this AI model.

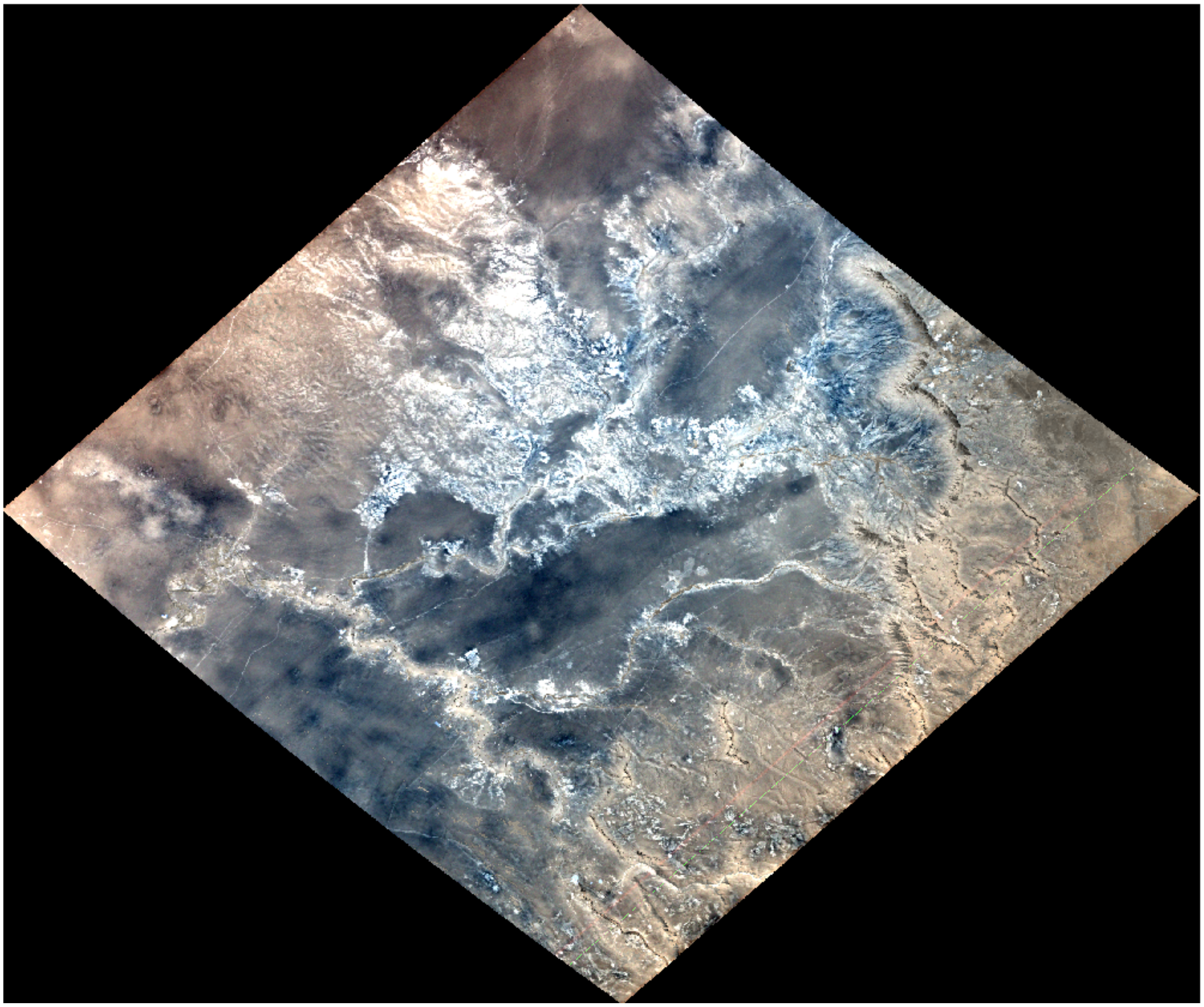

Figura 3. ENVI of a granule

This ENVI file can undergo various preprocessing techniques to reduce noise, but we have chosen to skip this step.

The next step is to find the endmembers. For this, the AVHYAS extension provides several methods, and we have used ATGP.

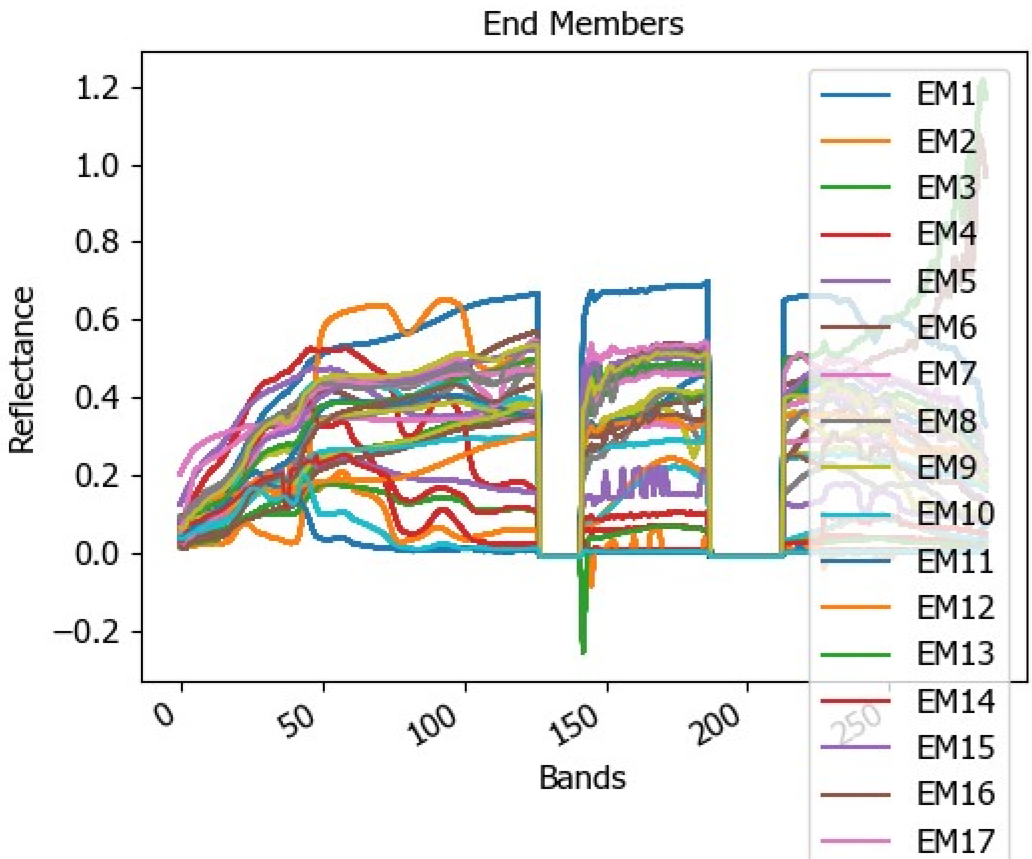

Figura 4. Endmembers of a granule

With the endmembers, we can generate an abundance map which will serve as a basis for creating a ground truth for the AI model.

However, this abundance map may not be very accurate as a ground truth, so we need to manually correct it.

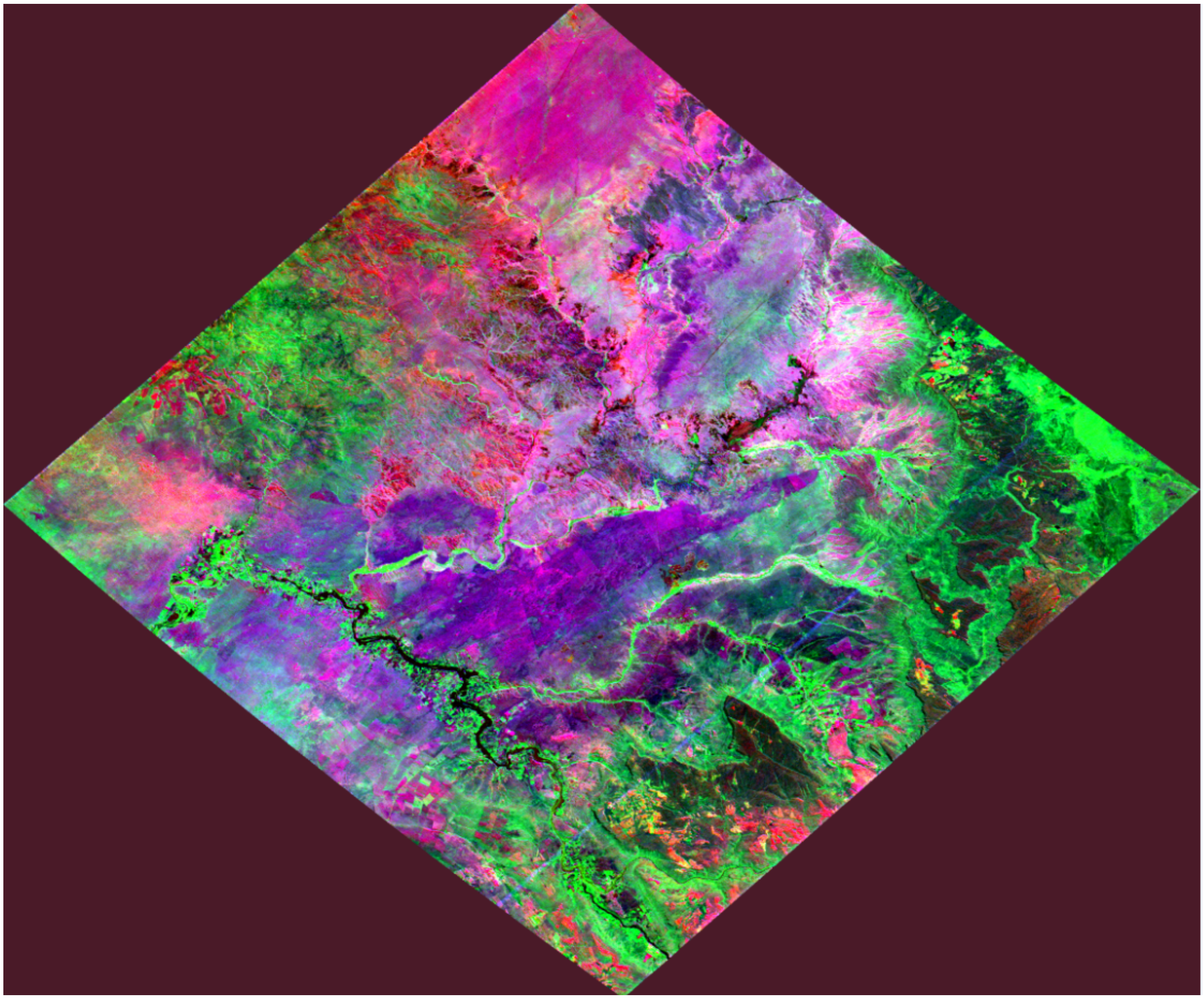

Figura 5. Abundance map de un gránulo

Once we have obtained an accurate ground truth, we can proceed to train our model.

The AVHYAS library provides us with various classification models. In this case, we have used Random Forest. When we apply this method to our raster, the result is a classified raster using the model. This model can be saved in Pickle format for later use in the data retrieval and processing algorithm.

6. Regards

Given the large memory footprint of satellite images and the physical limitations of our equipment, the algorithm outlined must progressively delete the images as they are processed to conserve resources. However, on more powerful systems, this aspect of the design can be overlooked.

Additionally, it's worth mentioning that a parallel version of the algorithm could be developed. This would allow multiple nodes to download and process images simultaneously, greatly reducing the time cost of the process.

7. Conclusions

Spectrometry is a highly complex field, with a wide variety of approaches, sensors, and different types of images to study. In this study, only a scaled-down version has been presented to illustrate how data would be obtained for later loading in the website demo.

In the future, both resources and the depth of the analysis could be expanded to provide a greater amount of information from each image, as well as to enhance the quality of visualization and accuracy of representation.