ANALYSIS OF VECTOR AND RASTER DATA USING OPEN SOURCE TECHNOLOGY

Branko Bogunović

Table of Contents

1. Data preprocessing

1.1. Input data

In this case study main objectives is to appropertly apply available techniques for analysing GIS data. In this example two different type of data was used (files are located in folder test_data):

1. Tabular data is data in table – points_ppi.csv.This file has three distinct columns:

   a) lon – longitude in degrees – WGS84 coordinate system (EPSG: 4326)

   b) lat – latitude in degrees – WGS84 coordinate system (EPSG: 4326)

   c) ppi – Plant Phenology Index (normalized data)

2. Raster data GeoTiff file - sentinel-2.tif – sentinel RGB image with resolution of 30m. This image is georeferenced into Croatian national coordinate system – HRTS- (EPSG: 3765).

1.2. Preprocessing data

One of the main steps to achive more objective data analysis is to examine all data before any further analysis is applied to the data. Therefore both data samples were examined and properly cleaned (filtered):

1.2.1. Tabular data

1. Tabular data were imported into Python. (read_data.py)
2. From insepection it was noticed data 12 data were missing in one of the column (lon, lat or ppi) Therefore, this data were removed from further analysis.
3. In addition to that PPI index were examined and there were 3 data samples were the values of PPI was above 1. In statistical analysis these three data samples (PPI > 1) would be categorized as outliers and therfore excluded from further analysis.

In addition to using Python scripts as a tool for preprocessing data, in folder /sql there is alternitive and more intuative way of importing, filtering data into database (PostgreSql). Databases are more suitable and optimized for tabular data. In this short sql script (multione.sql) tabular data were imported, filtered as well georefrenced into CRS:3765. In addition to that, database are working more efficient with large tabluar data (vector layers) rather then using solely Python.

After complete examination it was found, that 15 data samples were removed from further analysis which is around 1% of all data. This is not significant amount of data compared to original that would have impact on final results.

1.2.2. Raster data

Remote sensing is a vital technology for capturing data from a distance and provides wealth of information fo varoius fields like agriculture, ecology and urban planning. Howerver, raw images (Figure 1) often contain imperfecctions due to noise, limited contrast . On of the key techniques for removing noise is image enhancement techniques. Contrast augmentation is vital for enhancing the quality of image. One of the main fetaure of this techniques is enhances the perception of crucial elements in the image and therefore extracting useful data from the image.

Figure 1 shows, original raw image which is relatively dark and hard to see any specific features. Therefore, additional information from the image was obtained. In this example histogram of the image was used, that could provide us with further evidence how can image be enhanced. A histogram is graphs that show the statistical frequency of data distribution within a data set. In RGB image, color is represented as 3-digit(red, green blue) with values from 0-255.

In the Fig.2 it can be observed, that most of the range(for all 3 bands) is situated around 0. This is reason that image on Fig.1 is so dark and therefore it is hard to distinguish any particular feature.

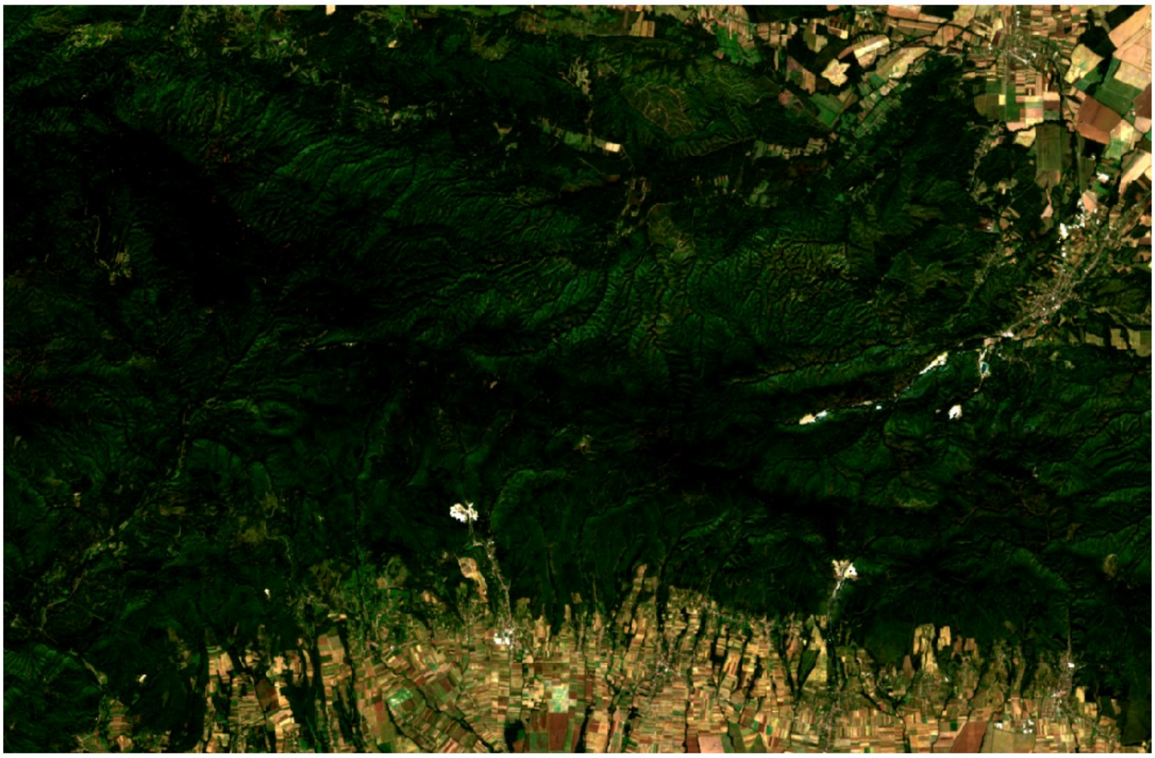

Figure 1: Original RGB raster (sentinel) layer

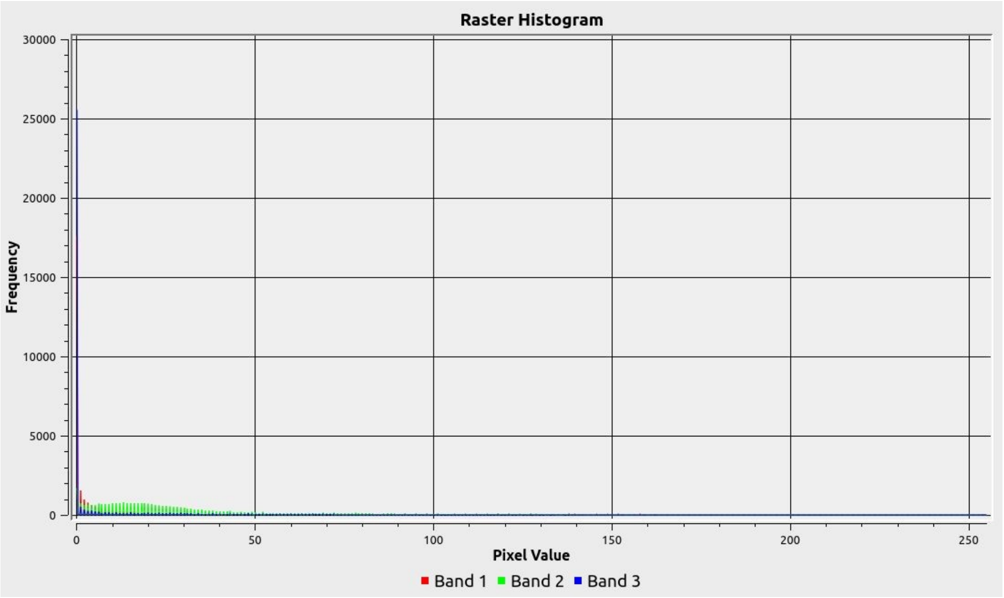

Figure 2: Histogram of original RGB image.

For this reason original image was enhanced with using CLAHE (Contrast Limited Adaptive Histogram Equalization). CLAHE is an adaptive method that enhances contrat while preventing over amplification of noise. Complete proccedure is explained in Python script contrast_enhac.py. On the Fig.3 is original image with application of CLAHE method. It can be seen that image quality improved significantly, comparing with original (Fig.1).

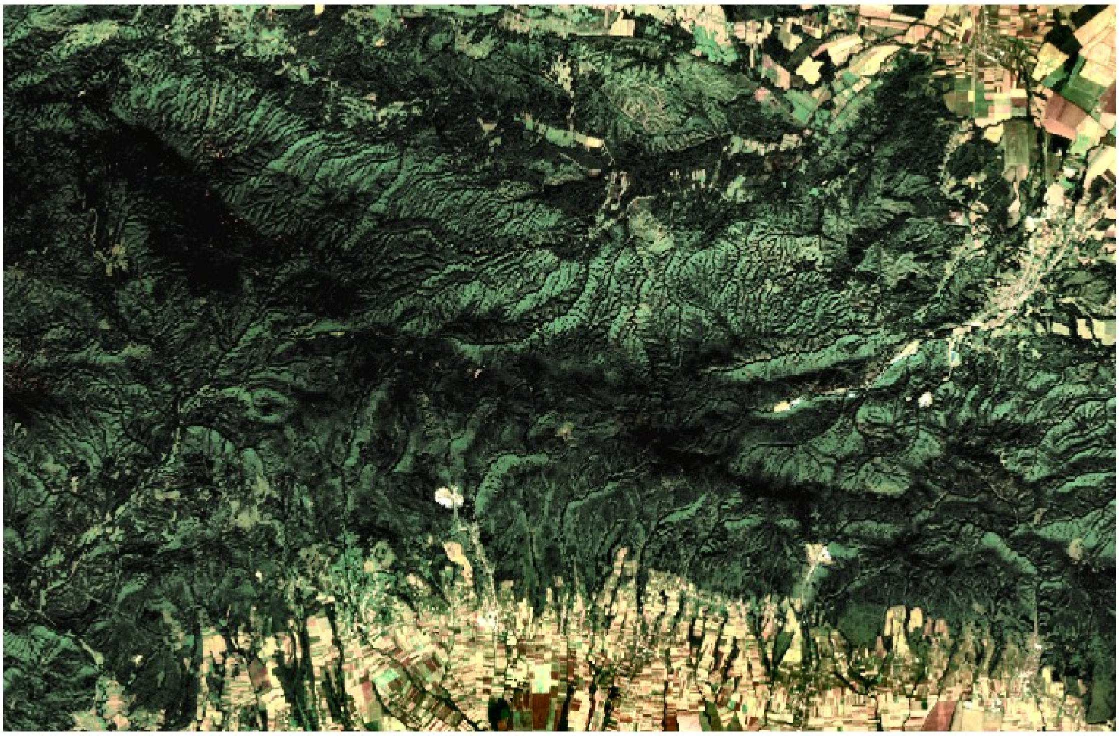

Figure 3: RGB image raster after using contarst enhancement

In addition to that, histogram from enhanced RGB image (Fig 3.) was produced to observe if there is any significant change in color distribution. Frequency for all 3 color bands are more evenly distributed and therefore quality of the image (Fig.3) improved significantly and subsequently features on the image are more observable .

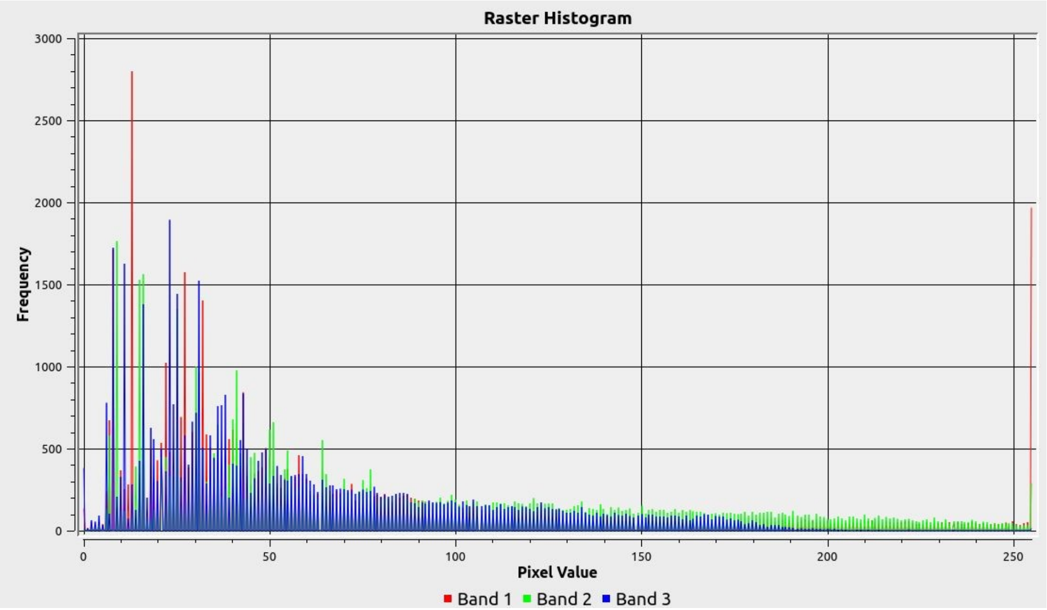

Figure 4: Histogram of filter RGB image after using CLAHE method.

2. Data analysis

2.1. Vector data

In this chapter geospatial analysis will be applied on tabular data. As explained in prevoius chapter data were georeferenced to appropriate coordinate system. With this we can overlay different type of data (vector and raster) and perform some advanced analysis.

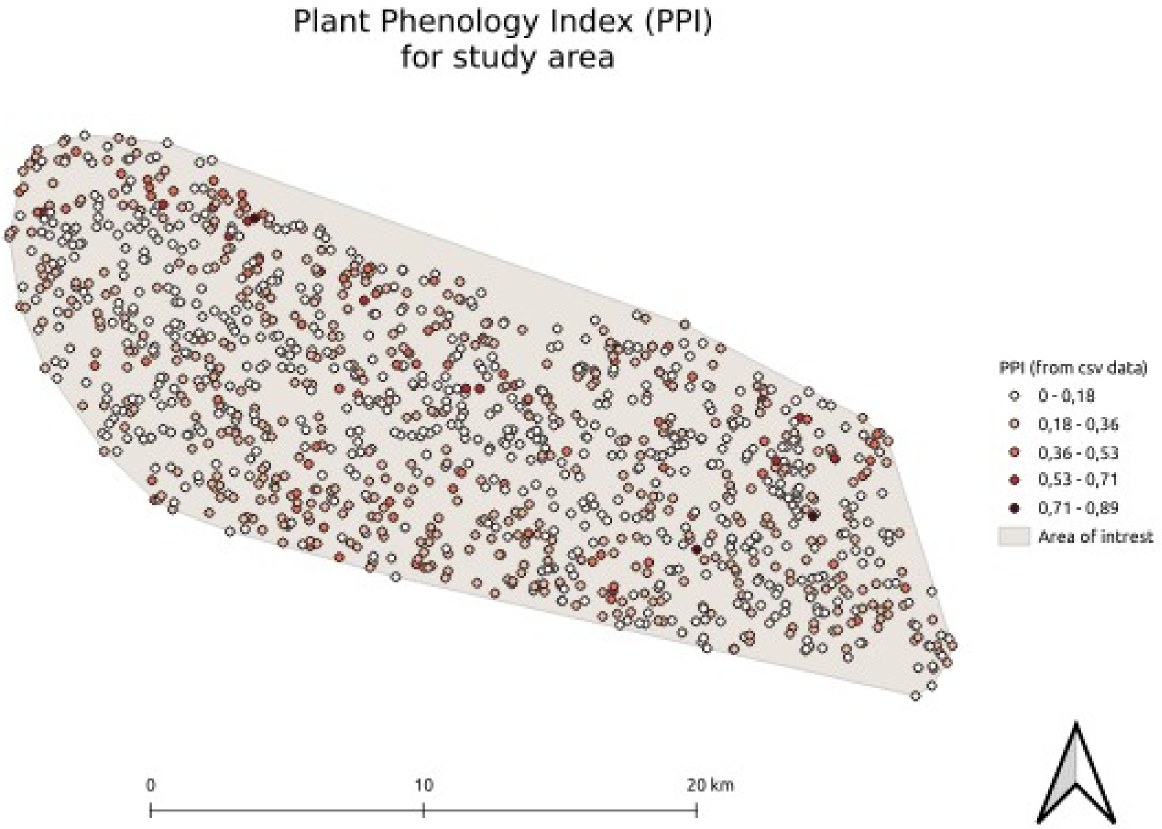

Figure 5: Spatial representation of PPI for study area using QGIS.

Notice from the Fig.5, that most of plant phenology index (PPI) for study area is situated between 0-0.36. On another hand high values of PPI is scares and found only on few locations. The shadow map present the border of the study area which was obtained through convex_hull function.

From the Fig.5 we can only identify how PPI is changing from one point to another. Furthermore, it is hard to distingusih how PPI is changing from one location to another through the space.

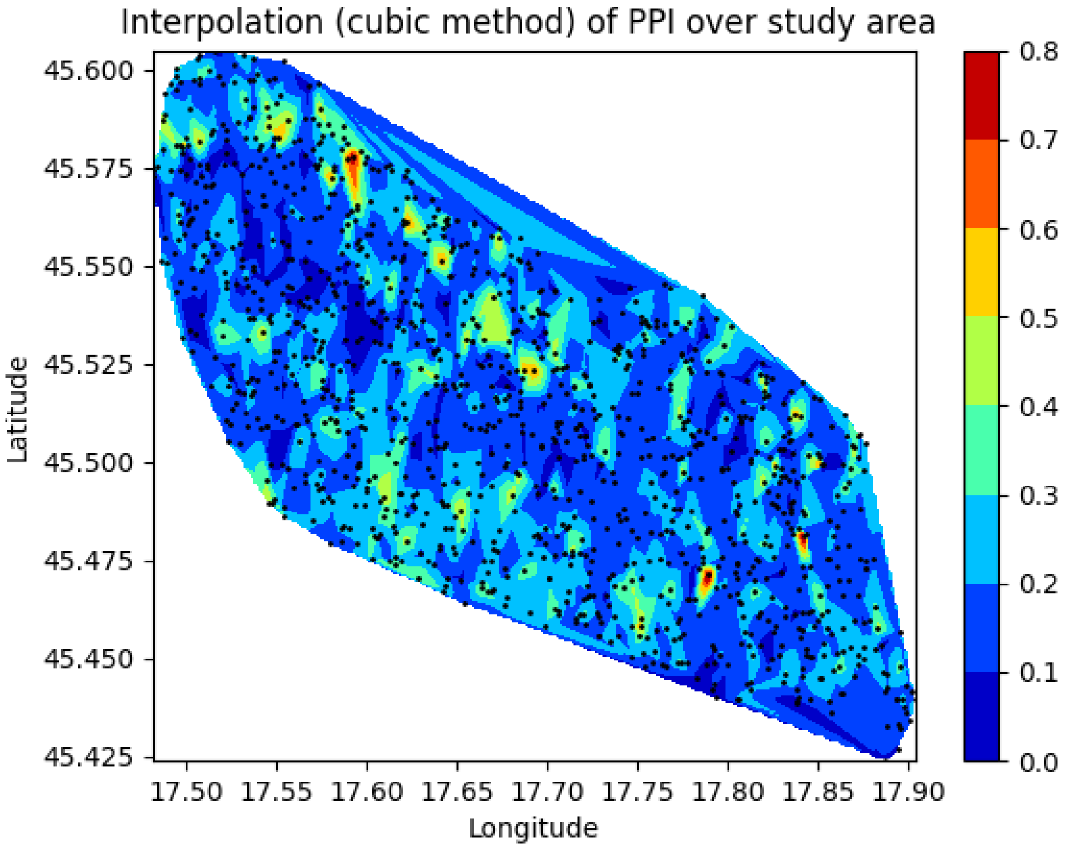

Figure 6. Spatial interpolation of the PPI for study area

To overcome this issue, interpolation of the PPI points were performed for the study area (*interpolation_2d.py*). In this example we used basic cubic spline method for interpolation. Fig.6. show spatial changes PPI and most area is covered with PPI < 0.3. There are only few area (3 locations) where PPI > 0.6.

This is one example how we can convert set of the points into polygon map. This is practical techniques suitable for further analysis particularly where combination of raster and vector data is needed.

2.2. Raster data

One of the fundemental aspect of remote sensing is image analysis. With appropriate analysis researcher can extract valuable information from images captured by satellites and other sensos. This chapther focuses on how can we implement some basic band math, indicies.
In our example will be using some basic indicies calculation from RGB sentinel image.
There are different type of indicies such as NDVI (Normalized Difference Vegetation Index) that can provide useful information about health of the plants. For calculation of the NDVI images, NIR (near infra red) and red bands has to be provided.
However, in our example only RGB (red, green and blue) image is provided and therefore we can found another approach to calculate indicies. In recent decades, researchers developed few indicies that are only based on RGB bands.
In this case study we evaluated three different type of the indicies (*ndi_calculation.py*):

a) Normalized Difference Index(1):

$$NDI = 128\times((G-R)/(G+R))+1$$

b) Normalized Green-Red Difference Index(2):

$$NGRDI = (G-R)/G+R$$

c) Green Leaf Index(3):

$$GLI = (2\times G-R-B) / (2\times G+R+B)$$

where, R - red band, G – green band and B – blue band.

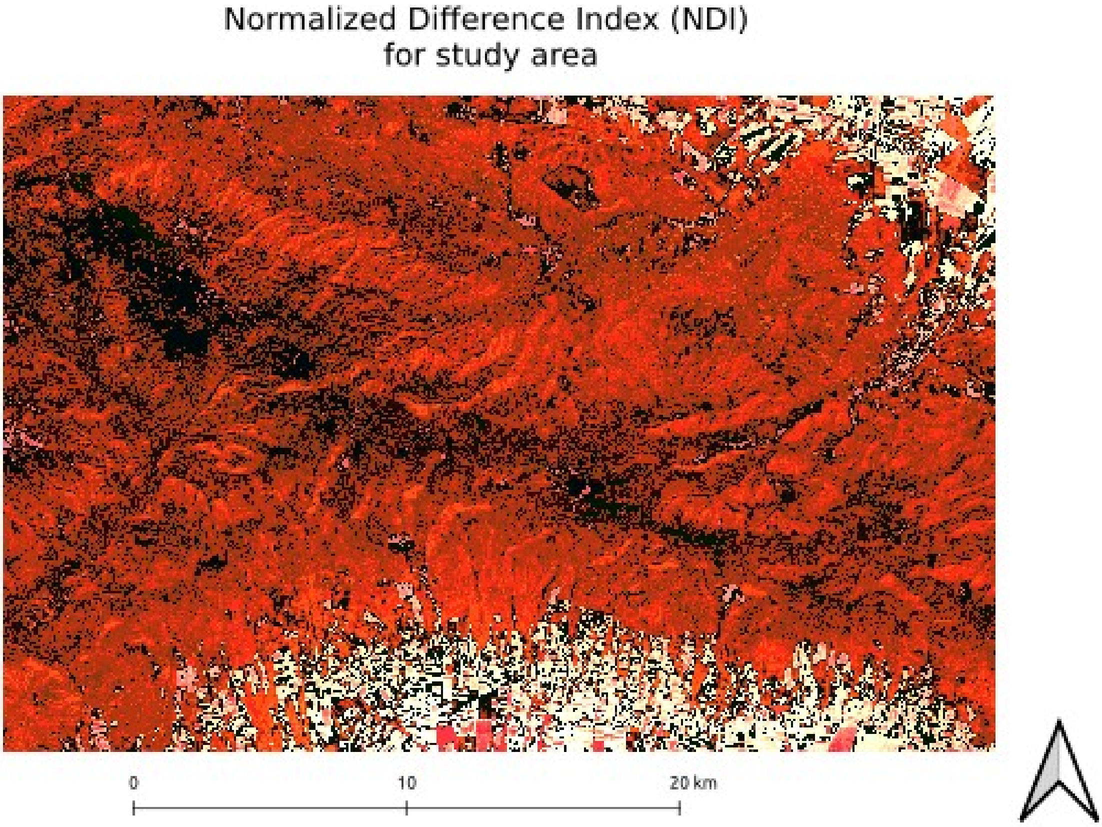

Figure 6. Normalized Difference Index (NDI

On the Fig.6. spatial representation of NDI for study is shown. This indices is helpful to meassure crop growth status.

2.3. Vector and raster data analysis

After analyzing each data (vector and raster) type seperatly, in some instances it is necessary to combine both data types. This is particularly true in image classification. Classification is essential method in remote sensing that allows for the grouping pixels (raster data) in an image to certain classes and categories, that could be provided from other sorces (vector data). This procedure is essential for task such as mapping land cover or identifying features.

Therefore it is important to find if there is any correlation between data. With scatter plot of NDI vs PPI it can be shown if there is any type of function which could be fitted (Fig.7).

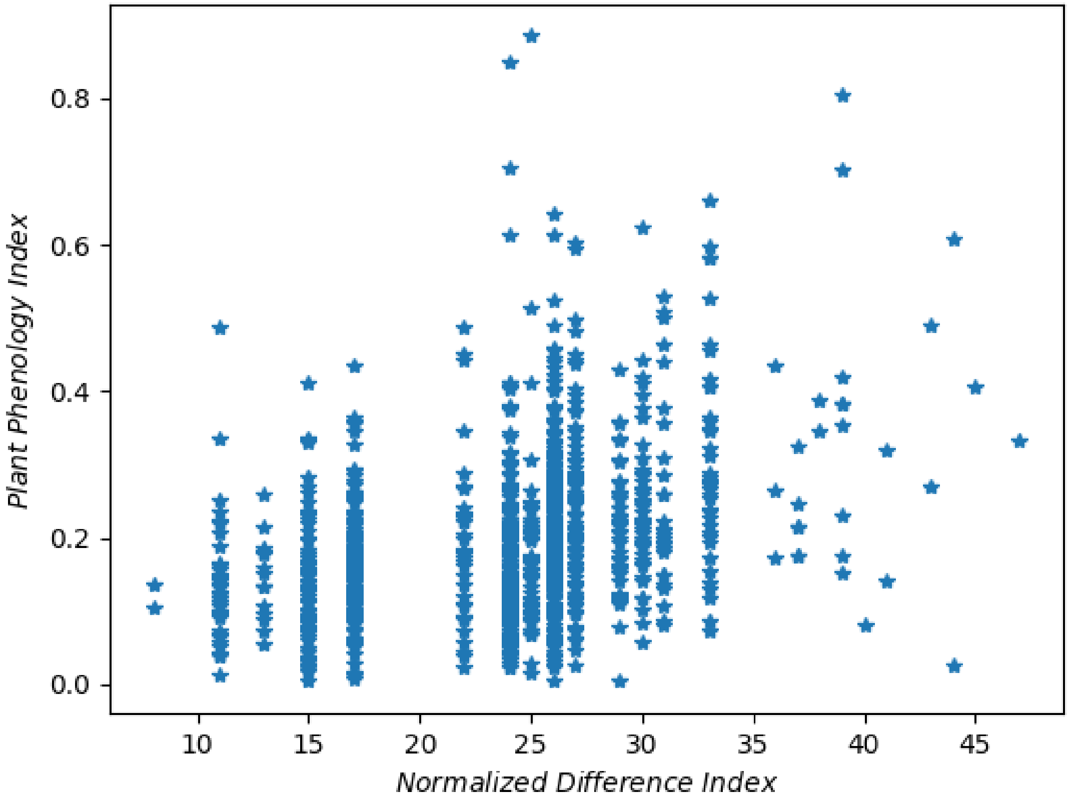

Figure 7: Scatter plot of NDI vs PPI

From the Fig.7 it can be noticed that it would be difficult to find any meangiful nonlinear regression model (ie. polynomial). By applying any classical non linear regression model would produce complex and overfitted model that would not represent any meangful model for training new data. Therefore it is more appropriate to use supervised machine learning alghoritm. Supervised machine learning alorithm can be trained to classify data into multiclass categories. In this case we decided to use SVM - support vector machine(*svm.py*). SVM is a powerful algorithm for classification which also finds the best boundary between different classes.

Before we have to apply SVM, PPI data were normalized into the 5 classes

1. Class --> PPI <= 0.2
2. Class --> PPI > 0.2 AND PPI <= 0.4
3. Class --> PPI > 0.4 AND PPI <= 0.6
4. Class --> PPI > 0.6 AND PPI <= 0.8
5. Class --> PPI > 0.2

After we generate set of training(30%) and testing data (70%) and by applying the radial basis function (RBF) kernel, we obtained confusion matrix.

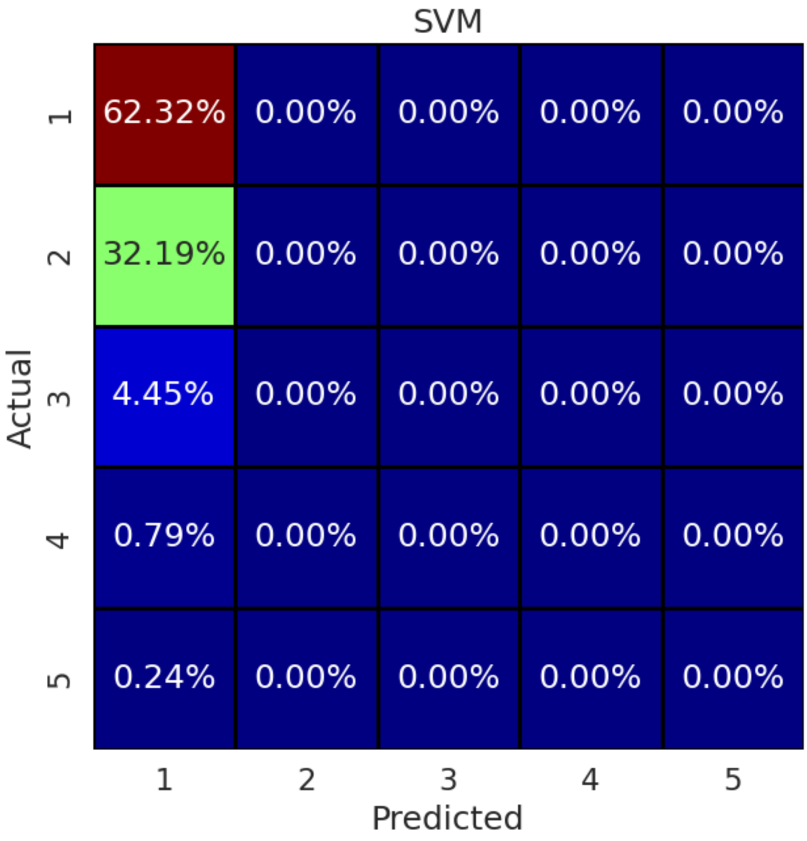

Figure 7: Confusion matrix for SVM method.

SVM prediction model accuracy was around 62% - for 1.Class (PPI values below 0.2). When we compare other other four classes 2-5 (PPI > 0.2) from confusion matrix, we notice that are predicted in 1.class. Reason for biased results could be in the nature of the PPI data, which are biased towrads 0.1 value.

To improve SVM prediction, more evenly based PPI data would be beneficial.

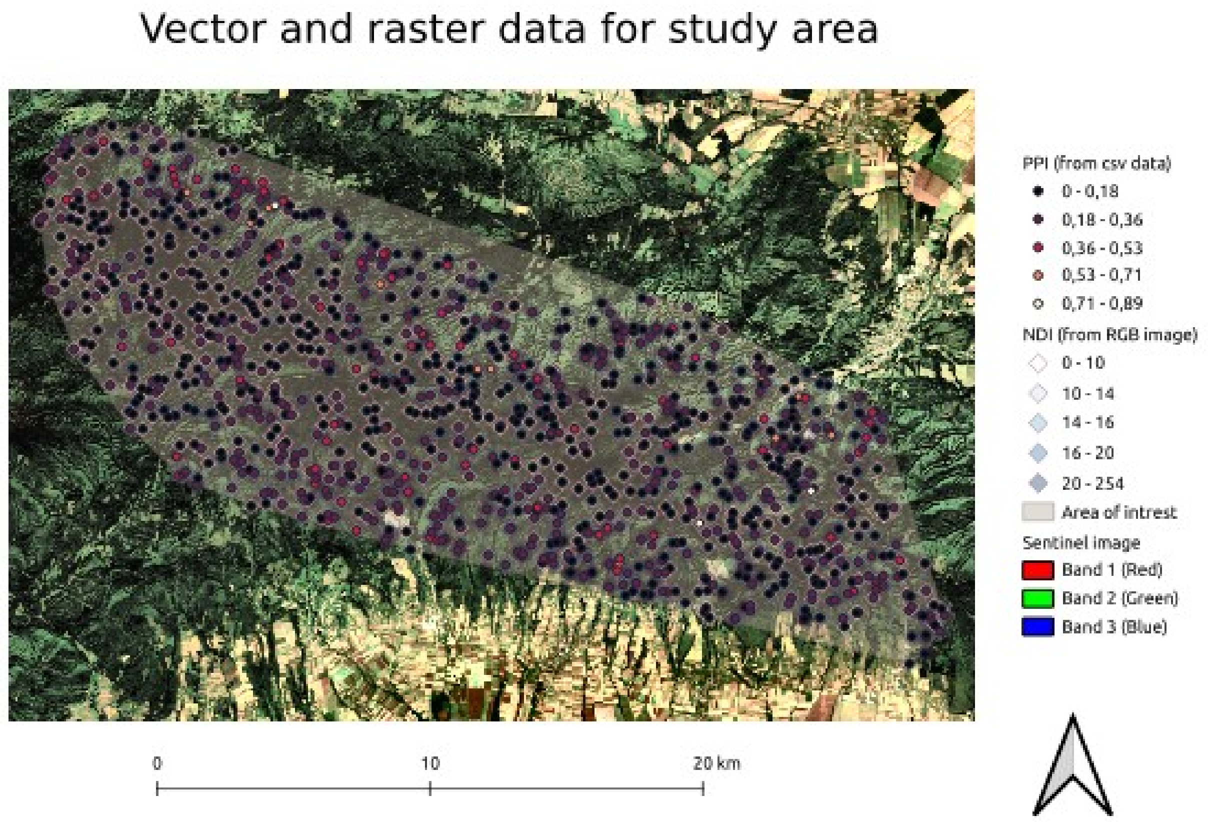

Figure 8: Spatial represtation of vector and raster data

3. Conclusion

Main objective of this case study was to show:

1. How to import, filter, geoproecessed data with Python. Furthermore, simple sql script was added, for importing tabular data into database.
2. Using different techniques for vector (interpolation) and raster (calculation of indicies)
3. Using machine learning techniques for classification of the RGB image.

This case study was build on very limited amount of data. With more data (NIR image, additional vector data) some more advanced analysis could be applied.

4. Literature:

1. Sonnentag, Oliver, Koen Hufkens, Cory Teshera-Sterne, Adam M. Young, Mark Friedl, Bobby H. Braswell, Thomas Milliman, John O’Keefe, and Andrew D. Richardson. "Digital repeat photography for phenological research in forest ecosystems." Agricultural and Forest Meteorology 152 (2012): 159-177. https://doi.org/10.1016/j.agrformet.2011.09.009

2. Perez, A. J., F. Lopez, J. V. Benlloch, and Svend Christensen. "Colour and shape analysis techniques for weed detection in cereal fields." Computers and electronics in agriculture 25, no. 3 (2000): 197-212. https://doi.org/10.1016/S0168-1699(99)00068-X

3. Hunt, E. Raymond, Michel Cavigelli, Craig ST Daughtry, James E. Mcmurtrey, and Charles L. Walthall. "Evaluation of digital photography from model aircraft for remote sensing of crop biomass and nitrogen status." Precision Agriculture 6, no. 4 (2005): 359-378. https://doi.org/10.1007/s11119-005-2324-5