**Solution report**

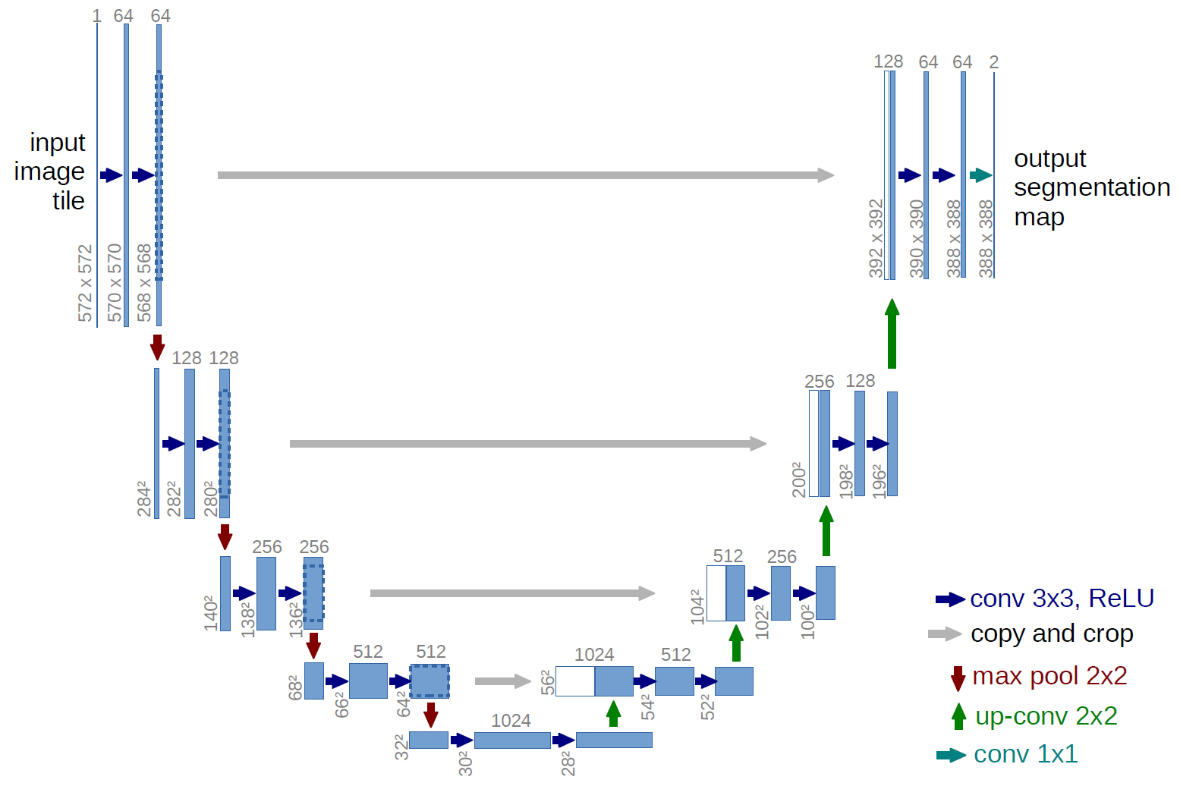
The first part of the tasks was made with a help of the link to the guide how to analysis the geo­data. <https://medium.datadriveninvestor.com/preparing-aerial-imagery-for-crop-classification-ce05d3601c68>

Firstly, I download QGIS to see the image in real size and to understand more about geo­data. Next step was to install Rasterio and geopandas libraries. With a help of these libraries I opened .jp2 file and mask file .shp.

Secondly, I tried to mask the image using rasterio.mask.mask() but no Rasterio failed to mask 937 files it is almost all masks that this file contain. So with the x and y values from the ‘geometry’ data and website <http://projfinder.com/> I found that the best epsg is 4200 Pulkovo. First time I tried another epsg and even find the part of the field in google maps and I notice that the coordinates on map doesn’t fit so correct as I had in x and y values. So, when I entered new values that I found I got Pulkovsk epsg that can recognize 1 more mask. As a result, the number of masks reduced to 434. I think it is still a lot but I couldn’t find better solution then changing epsg.

Thirdly, I prepared data for semantic segmentation. I made a mask that contains every erosion detection that were captured and split this image into small images 512x512. I needed this size for neural network called U-Net because the input size of this network is an image of this size. Moreover, I spit the .jp2 image the same way. So, I got 484 images for training the model. I made two folders for train and test in which there are also two folders for input images and output binary masks.

The second part of the task I made a neural network that I used before for the segmentation. This network is called U-Net. The model of this network we can see on the picture 1



Picture 1 – U-Net

There are different layers in this network such as: Input, Convolution, Maxpooling, Dropout and concatenate layers. Convolution layers are used extract features from the image. We try to teach the net to identify the erosion by itself so it needs to extract the features that after we can com­pare with binary image. Maxpooling we need not only for decrease the image but also for getting the maximum value of filter. The output after max-pooling layer would be a feature map contain­ing the most prominent features of the previous feature map. It is used Dropout to prevent over­fitting by turning of some neurons. Concatenative layers used to connect two different layers to one.

As a result after model.summary() we got total params: 1,941,105 it is big amount of parameters and the model we be training for a long time. I made checkpoint to save the best model that was trained. I also wrote callbacks to prevent overfitting so the model is observing the validation loss. If loss will grow more then two times, the training process will stop and save the result it got 2 steps before. The reason why the model check validation loss is because when it is getting higher the model starts to overfit. That’s why it is observed. After training I predict new values using test data in the folder test. However, as I wrote in the REDME file I think that, due to the fact that I read not every erosion from the mask file, the final binary mask contains a lot of empty space where there is nothing to segment. So, these files are much more then files that contains erosion. That’s why after predicting new values we got just empty binary image. The way to improve this model is to read data more accurate that I tried to do the last days however, no results. The U-Net model should work correct because it was tested with another data and showed high accuracy. I wish to finally understand how to work with geodata correctly and redo the analysis part.

**Solve problem about erosion detection**

The solution was made by this research <https://www.mdpi.com/2072-4292/11/5/513/htm#B5-remotesensing-11-00513>

One of the factors of soil erosion is a height where we can observe it. Zhang et al. [**[5](https://www.mdpi.com/2072-4292/11/5/513/htm" \l "B5-remotesensing-11-00513)**] assessed soil erosion from remote sensing data in rehabilitated high-density forests of Hetian, a typical red soil region in southern China and found that the eroded areas were primarily distributed in locations with elevations between 300 to 500 m.( In addition, some areas with elevation between 300 to 500 m and slope angles below 25°, primarily in the southeastern and southwestern parts of the study area, also continue to suffer from soil erosion.) We can find height using epipolar geometry.

There are a number of factors that influence soil erosion and are therefore involved in model construction for soil erosion detection. Of them, rainfall is one of the frequently-considered factors as rainfall-induced soil erosion occurs commonly in various environments. Long-term rainfall data are useful to identify the period of strong soil erosion, when used as the rainfall erosivity factor (R-factor) of RUSLE for investigating rainfall erosivity variability [[**37**](https://www.mdpi.com/2072-4292/11/5/513/htm#B37-remotesensing-11-00513)]. **Rainfall influences land surfaces by its duration**, intensity and cumulative amount per event. A recent study in a clayey coastal hilly area in Italy showed that soil erosion was correlated more with cumulative rainfall than with peak rainfall intensity (<https://www.mdpi.com/2073-4441/10/10/1314/htm>)

Slope is another remote sensing data derivable factor that has been commonly-used in soil erosion monitoring as many of deriving forces to soil erosion have been associated with slope [[**9**](https://www.mdpi.com/2072-4292/11/5/513/htm#B9-remotesensing-11-00513),[**11**](https://www.mdpi.com/2072-4292/11/5/513/htm#B11-remotesensing-11-00513),[**15**](https://www.mdpi.com/2072-4292/11/5/513/htm#B15-remotesensing-11-00513),[**49**](https://www.mdpi.com/2072-4292/11/5/513/htm#B49-remotesensing-11-00513),[**50**](https://www.mdpi.com/2072-4292/11/5/513/htm#B50-remotesensing-11-00513)]. Slope is an important basic input of many soil erosion models currently in operation in the world, such as those of [[**9**](https://www.mdpi.com/2072-4292/11/5/513/htm#B9-remotesensing-11-00513),[**11**](https://www.mdpi.com/2072-4292/11/5/513/htm#B11-remotesensing-11-00513),[**14**](https://www.mdpi.com/2072-4292/11/5/513/htm#B14-remotesensing-11-00513),[**15**](https://www.mdpi.com/2072-4292/11/5/513/htm#B15-remotesensing-11-00513),[**17**](https://www.mdpi.com/2072-4292/11/5/513/htm#B17-remotesensing-11-00513)]. The factor can be represented by slope steepness and slope length [[**9**](https://www.mdpi.com/2072-4292/11/5/513/htm#B9-remotesensing-11-00513),[**11**](https://www.mdpi.com/2072-4292/11/5/513/htm#B11-remotesensing-11-00513),[**17**](https://www.mdpi.com/2072-4292/11/5/513/htm#B17-remotesensing-11-00513)], while the latter is not as popular as the former. **A recent** study in the karst regions of southwestern China **found** that **slope degree influenced** soil loss **more than** **slope length** [[**51**](https://www.mdpi.com/2072-4292/11/5/513/htm#B51-remotesensing-11-00513)].

Healthy trees reduce raindrop-induced soil erosion by forming a dense and multi-storey canopy with plenty of branches and leaves and, therefore, can greatly dissipate the kinetic energy of raindrops before reaching the ground surface. In addition, healthy trees often have more developed root zones, which reduce soil erosion potential. In contrast, poor healthy trees with sparse branches and leaves cannot effectively obstruct the strike of raindrops and hence are unable to reduce the rainfall erosivity and fails to protect the soil from the direct impact of raindrops and throughfall [[**58**](https://www.mdpi.com/2072-4292/11/5/513/htm#B58-remotesensing-11-00513)]. Yan et al. [[**59**](https://www.mdpi.com/2072-4292/11/5/513/htm#B59-remotesensing-11-00513)] and Zhu et al. [[**4**](https://www.mdpi.com/2072-4292/11/5/513/htm#B4-remotesensing-11-00513)] reported that the areas covered with unhealthy trees tended to be more likely to develop soil erosion.

Accordingly, the major factors that influence soil erosion in forest are FVC, vegetation health status, soil exposure degree and slope. The vegetation health status can be determined by two factors, yellow leaf index and nitrogen index. The selected factors are essential to identify the spots in forest that are particularly prone to erosion and thus were considered in the SEUFM model development.

Carlson and Ripley:

*FVC* = [(*NDVI* −*NDVI*0)/(*NDVI*∞ − *NDVI*0)]2

(3)

Gutman and Ignatov:

*FVC* = (*NDVI* − *NDVI*0)/(*NDVI*∞ − *NDVI*0)

(4)

where *NDVI* is the *NDVI* value of a pixel, *NDVI*0 is the *NDVI* value for bare soil selected from entirely bare soil and *NDVI*∞ corresponds to the *NDVI* value of a surface with an *FVC* of 100%, selected from extremely dense forest. The soil erosion under forest model (SEUFM) was developed with the factors including FVC, NRI, YLI, NDSI and slope. Due to the difference in unit and data ranges, the five factors have to be normalized within [0, 1] before they can be integrated.

The five normalized factors were integrated to form SEUFM using two methods, that is, principal components analysis (PCA) and the multiplication approach.

PCA is a statistical technique that converts a set of measurements into a set of values of uncorrelated principal components (PCs) using an orthogonal transformation. PCA automatically weighs the contribution of each variable into each principal component based on the variable’s loading [[**69**](https://www.mdpi.com/2072-4292/11/5/513/htm#B69-remotesensing-11-00513)]. The integration of the selected five factors in a model involves an evaluation of the contribution of each factor to soil erosion, which is very likely to be different. This suggests that a weight may have to be assigned to each factor. If the soil erosion can be associated with one or more PCs, PCA can be a good method to integrate the selected factors for SEUFM because PCA can automatically quantify the contribution of each factor into each PC according to factor loadings

Soil erosion in forests is closely related to, and can be reflected in, factors such as forest coverage, forest health status, soil exposure intensity and slope. These are the factors considered in this study for the development of the SEUFMs. Of the two types of SEUFM models developed in this study, the PCA-based method achieved higher accuracy than the multiplication-based method because it handles the five factors according to their contributions (loadings) to each PC.