QGIS AND ITS OPTIONS OF IMAGE CLASSIFICATIONS

Content

BACKGROUND OF IMAGE CLASSIFICATIONS

Image classification is the process of **assigning classes to different pixels or groups of image pixels (objects)**, i.e. summarizing the image to the categories. The basic unit of the image is a pixel (one cell), classification tool works with the (reflectance) statistics of individual pixels. Classification can classify the total scene content into a limited number of classes or distinguish one or more specific classes within the landscape.

Image analysis can be performed on multispectral as well as hyperspectral imagery or any other imagery. As numerous factors affect the results, the image classification is rarely perfect. Important factors are the spectral and spatial characteristics of the data, the natural variability of terrain conditions in the geographic region, or the digital classification technique employed etc. The classification requires preparatory processing and the refinement of classes after the classification process with the use of ancillary data (e.g. terrain elevation and other parameters that influence land cover). In some cases, it may be useful to e.g. merge classes.

In general, there are two main image classification techniques – **pixel-based** and **object-based**. Both of them may be supervised by the user or not, i.e. **supervised** or **unsupervised** approach. Pixel-based image classification techniques are the most known and common approach; however, the object-based classification may be beneficial in the case of high-resolution data. Pixel-based approach assigns each pixel to a class; object-based approach assigns a group of pixels (objects) to a class. The essential step for classifying the image using the object-based approach is to segment the image into the objects (the process of segmentation should come before the classification alone).

Unsupervised classification is the most basic technique because you do not need samples. In unsupervised classification, it first groups pixels into "clusters" based on their properties using some algorithm such as K-means or ISODATA. After picking a clustering algorithm, you identify the number of classes you want to create and manually identify each cluster with land cover type. The process is usually iterated several times to acquire more trustable results and also to obtain some information about the clustering quality (Figure 1). See the example of clustering iteration there.
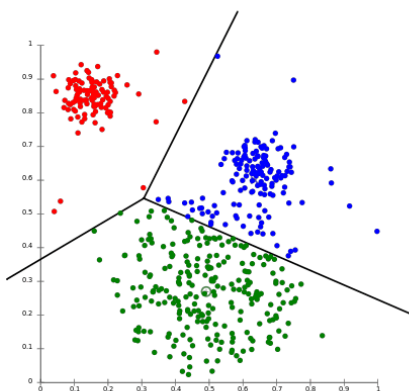

Figure 1. K-means data clustering during one of the iterations.

In supervised classification, you have to select representative samples for class (i.e. train your classification). The software then uses these "training sites" and applies them to the entire image. Supervised classification uses the spectral signature defined in the training set. The common supervised classification algorithms are maximum likelihood and minimum-distance algorithms.

Multiresolution segmentation (object-based image analysis - OBIA) produces homogenous image objects by grouping pixels. It generates objects with different scales. These objects may be more meaningful because they represent whole features (not only the pixel) in the image.

Read the textbook for more information.

UNSUPERVISED PIXEL-BASED IMAGE CLASSIFICATION

1) Run QGIS, open SCP plugin and load a *Band set* (i.e. data you want to classify).

2) Open *Band processing* tab and open the *Clustering* option.

3) Choose K-means or ISODATA classifier, specify the numbers of classes and iterations. Run the tool (Figure 2). New raster file will be created. *Note: It takes some to finish the clustering, it's better to start with one band only and the limited number of iterations.*

The number of classes means the number of groups to which will be data clustered. Choose the number with regards to the number of categories you want to distinguish, use as many or as less as needed. More classes can better differentiate various objects better; however, it may also introduce errors. You can merge the several classes, but you can't split a class afterwards. The more classes you choose, the more time it will take. A maximum number of iterations value use as much as possible. Iteration stops after these many iterations even if the convergence criterion is not satisfied. The higher number of iterations increases the reliably of the results, however, also increase the computing time.
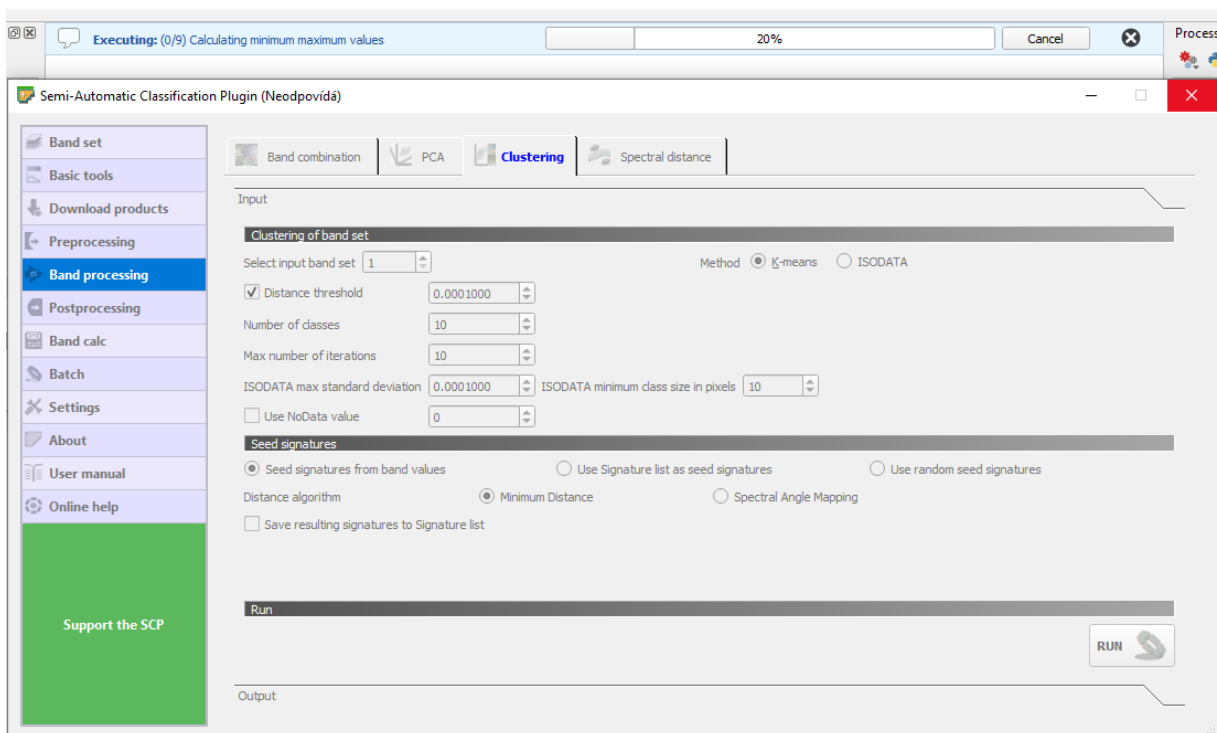

Figure 2. Options of unsupervised classification.

See Figure 3 for a comparison of K-means classification and true-colour composition. You may also check the output statistical report (*SCP – Band processing – Clustering – Output*)

Figure 3. Result of unsupervised classification (top) and true-colour image of the same place (bottom).

SUPERVISED PIXEL-BASED IMAGE CLASSIFICATION

1) Run QGIS, open SCP plugin and load a *Band set* (i.e. data you want to classify).

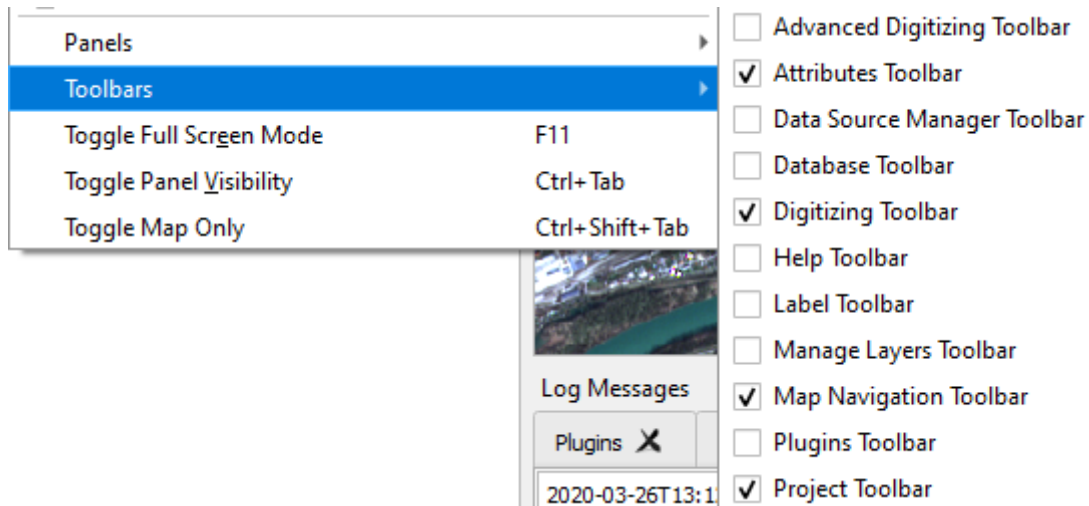**2)** Activate SCP toolbars, if needed (Menu *View – Toolbars*), see Figure 5.

Figure 4. Toolbars activation.

3) Open SCP Dock, if necessary, by clicking on the plus button (Figure 11).
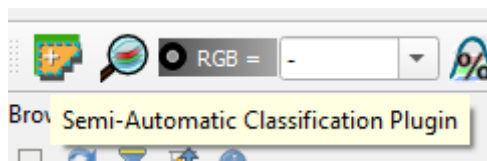

Figure 5. Opening SCP Dock.

4) Now you must create a file contain Regions of Interest (ROIs) to supervise your classification process. To do so, create a new training file at *Training input* tab (Figure 6). *Note: As usual, a Band set must be loaded*.
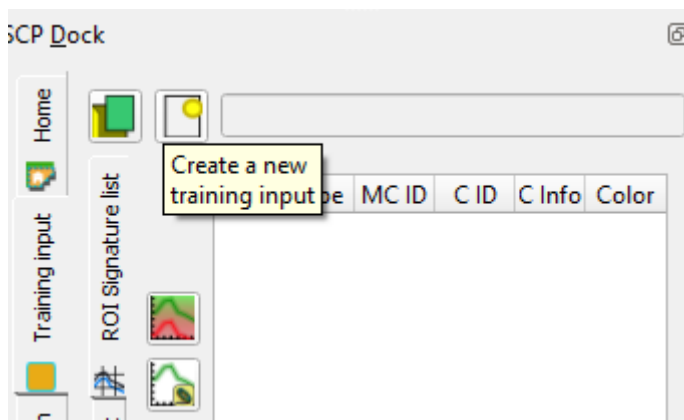

Figure 6. Creating of training file.

5) Once you have created a ROIs file, you can create specific regions, i.e. training polygons using *Create a ROI polygon* tool (Figure 7).
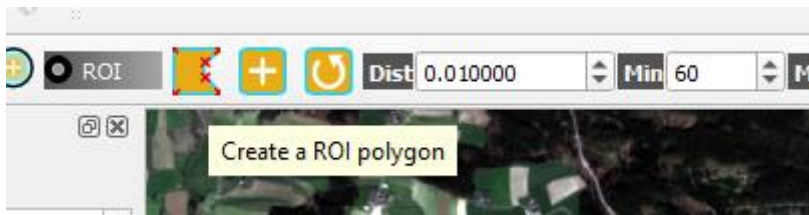

Figure 7. Creating a new ROI.

6) Using the left mouse button, you can create a polygon in the image, using the right mouse button you finish the sketch. Once you finish sketch you have to specify its parameters. Within the SCP plugin, it is possible to specify a Macro-class (MC) and Sub-class (C). This is a two-level differentiation of class level, Macro-class means the first/higher level, while Sub-class is the second (more detailed) level. For example, you can distinguish Water as a Macro-class and River, Lake and Pond as Sub-classes. Once you have created a polygon a specify its details, you can click on *Save temporary ROI to training input* (the most bottom right button), Figure 8.

Note: For each of intended (sub)classes you should create more than one ROI. The number of ROIs for each class is not given tightly, however, let's consider the number of five. Moreover, the ROIs for one (sub)class should be dislocated across the whole image. The ROIs for the same (sub)class very close to each other may introduce uncertainties.
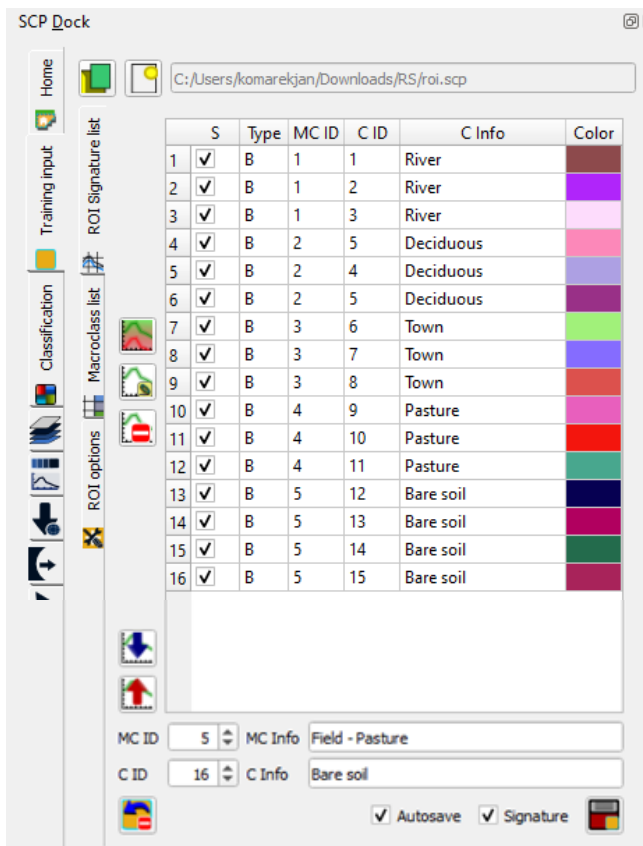

Figure 8. SCP Dock - Creating new ROI polygons for specific land cover classes.

7) Once you have created as many ROIs as you want, continue to the *Classification* tab at SCP Dock. There you can select the classifier and decide if Macro- or Sub-classes will be used (checkboxes at the top). You can check the options for creating a classification report as well as classified vectors (its optional), see Figure 9. Conversion of classification result into raster may be done also later using the *Postprocessing* tab in SCP.

Note: There are different classification algorithms: Minimum Distance, Maximum Likelihood or Spectral Angle Mapper. It always depends on the approach and the data which algorithm works the best. You should follow further reading as well.
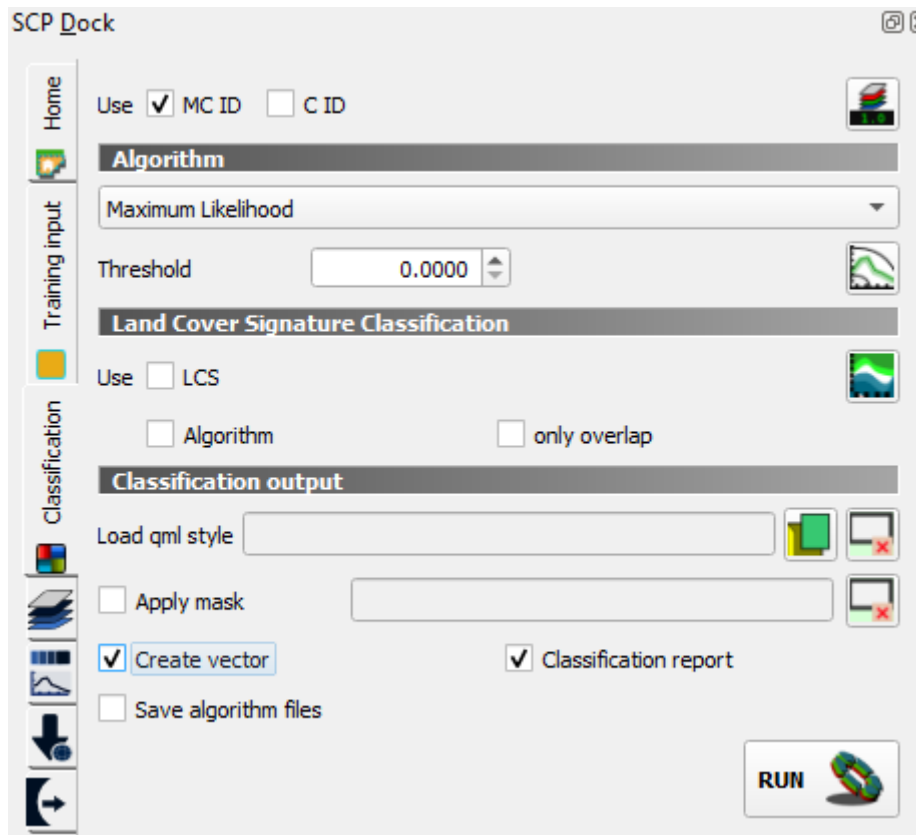
Figure 9. Options for supervised classification.

8) The result of the classification will appear at the Layer panel on the left, both raster and vector layers (if you select it as an output), Figure 10. Those layers together with the Classification report are stored at a specified folder. You can change layer symbology in the layer properties.
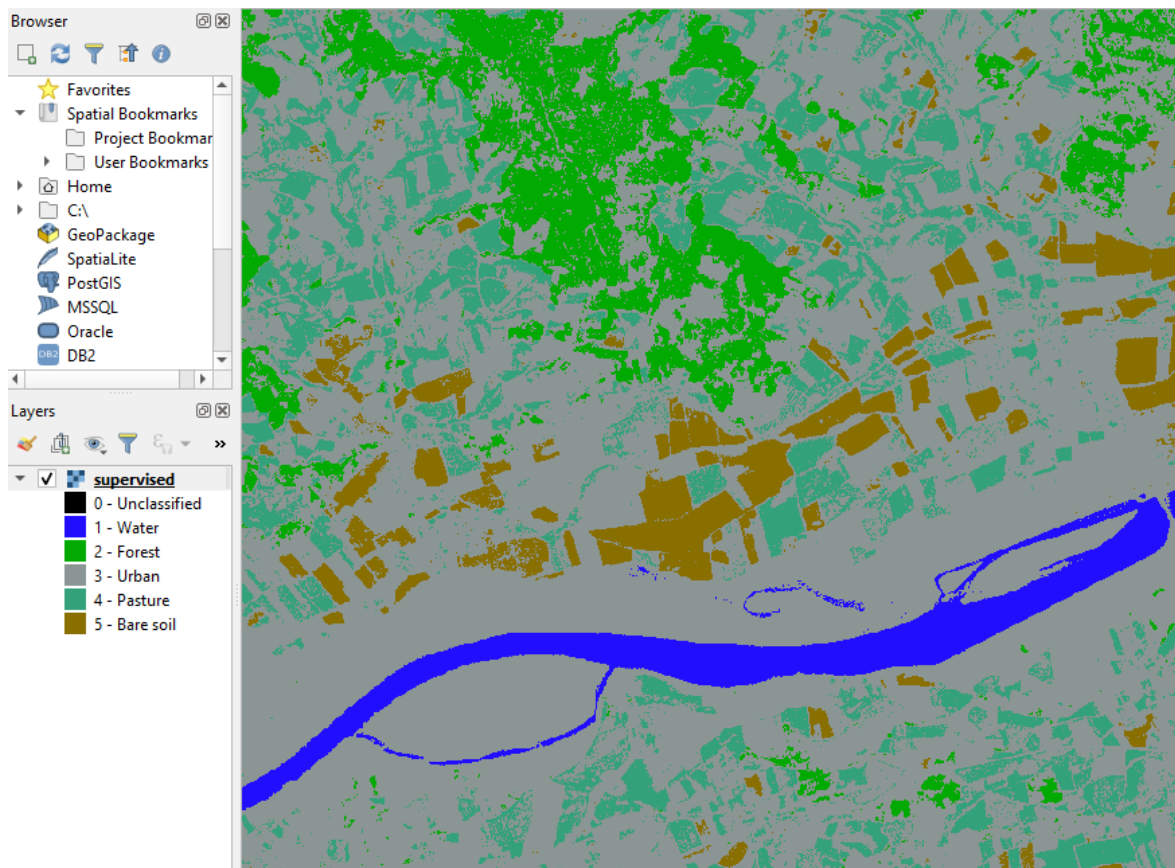
Figure 10. Supervised classification result.

CLASSIFICATION RESULTS POST-PROCESSING

Basic descriptive statistics

Once you classify the image, you should check statistical reports and try to quantify the quality of your results. You can generate a basic classification report for descriptive statistics about your pixel classification (*SCP –Postprocessing – Classification report– Output), see* Figure 11. Here you can read information about the total number of pixels assigned to each of the classes, its percentage ratio and area.
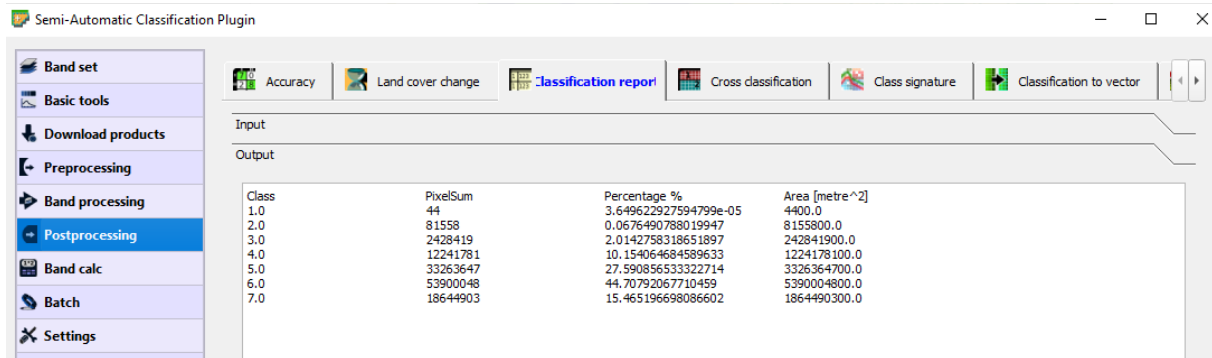
Figure 11. Classification report.

Subjective quality evaluation

The easiest option how to check the resulting quality is to check classes assignment comparing the classes with the true colour composition. Load the classification result above the true colour image and zoom to some area of the picture. Checking and unchecking the classification layer allows you to verify the classes. Probably at least some of the classes are misinterpreted with the reality. That's happened quite often because the spectral signature could be similar for different land cover types or objects. To check that, open the Spectral Signature Plot from the SCP Dock and load misclassified classes.

Look at my example below, Figure 12. Two classes are uploaded – urban (red) and bare soil (green). The range of the urban class is huge and practically contains the signature of bare soil. The reason is that there is a large spectral heterogeneity within the urban areas and could be easily misinterpreted with the surrounding land cover types (Figure 13). The solution for that may be in the specifying more ROIs for those classes or/and include some of the indices as input layers for the classification process.
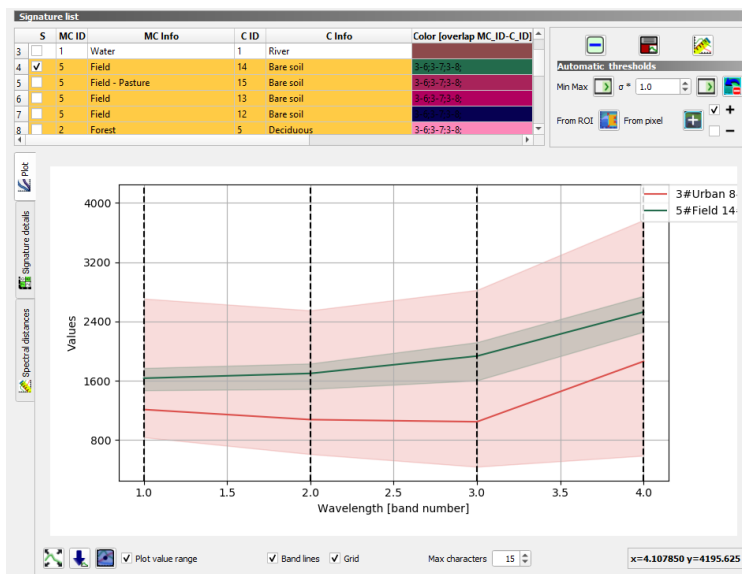
Figure 12. Spectral Signature Plot.

Figure 13. Spectral diversity across the urban areas.

Note: The SCP provides a lot of options to achieve a good classification result. It depends on the approach, how much time one wants to spend to improve the classification. The SCP provides even more options to improve the ROIs while altering the spectral signatures for different classes. Nonetheless, it will not be possible to classify every single pixel right.

Classification sieve

Your classification result may be somehow strewed, that's quite usual. Many isolated pixels may appear at the resulting image. It's beneficial to apply a sieve, which removes isolated pixels as well as may fill over isolated holes (Figure 14). You can use the *Classification sieve* tool at the *Postprocessing* tab within SCP. Two options are to be set *Size threshold* and *Image connection*. The threshold is the size of the patch to be replaced (in pixel unit); all patches smaller the selected number of pixels will be replaced by the value of the largest neighbour patch. Image connection is the number of considered pixels within the 3x3 window of focal statistics (i.e. 4, where diagonal pixels are not considered connected or 8, where diagonal pixels are considered connected.
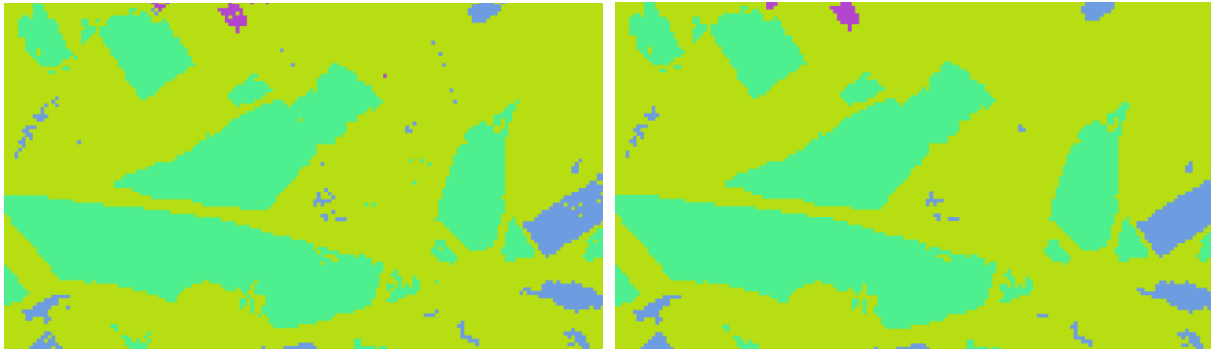
Figure 14. Classification sieve. Original classification result (left) and the sieved result (right).

Note: Other tools may be used to make the classification result better. In SCP at least, Classification erosion and Classification dilation are available. Classification erosion toll removes the border of a class patch (i.e. like erosion works), defining the class values to be eroded and the number of pixels from the border. Classification dilation makes the opposite; allows for dilating the border of a class patch, defining the class values to be dilated and the number of pixels from the border.

Quantitative assessment of the classification

To have an objective quantified it is necessary to use some of the statistical calculations, typically overall accuracy, user's accuracy, producer's accuracy, and Kappa. SCP allows to quantitative assess you results using the *Accuracy Assessment* tool at the *Postprocessing* tab. There your classification is compared to a reference raster or reference shapefile. The output is an error raster, where pixel values represent the categories of comparison (i.e. combinations identified by the error matrix) between the classification and reference. A text file containing the error matrix is created.

To do accuracy assessment you have to create a set of ROIs with the same classes as you used as training samples, however, to retrieve meaningful results you can't use the same ROIs as you used for the classification. Therefore, before you can run the tool, new ROIs have to be created the same way as it was done before. Once you have them, upload ROIs for validation to the *Accuracy Assessment* tool (Figure 15).

Note: The tool compares the correctly assigned classes using reference data (ROIs for validation). It is necessary to keep the same order of classes in both ROIs for training as well as for validation. If not, the results are meaningfulness (you, for example, may compare correctly classified pastures against urban areas).
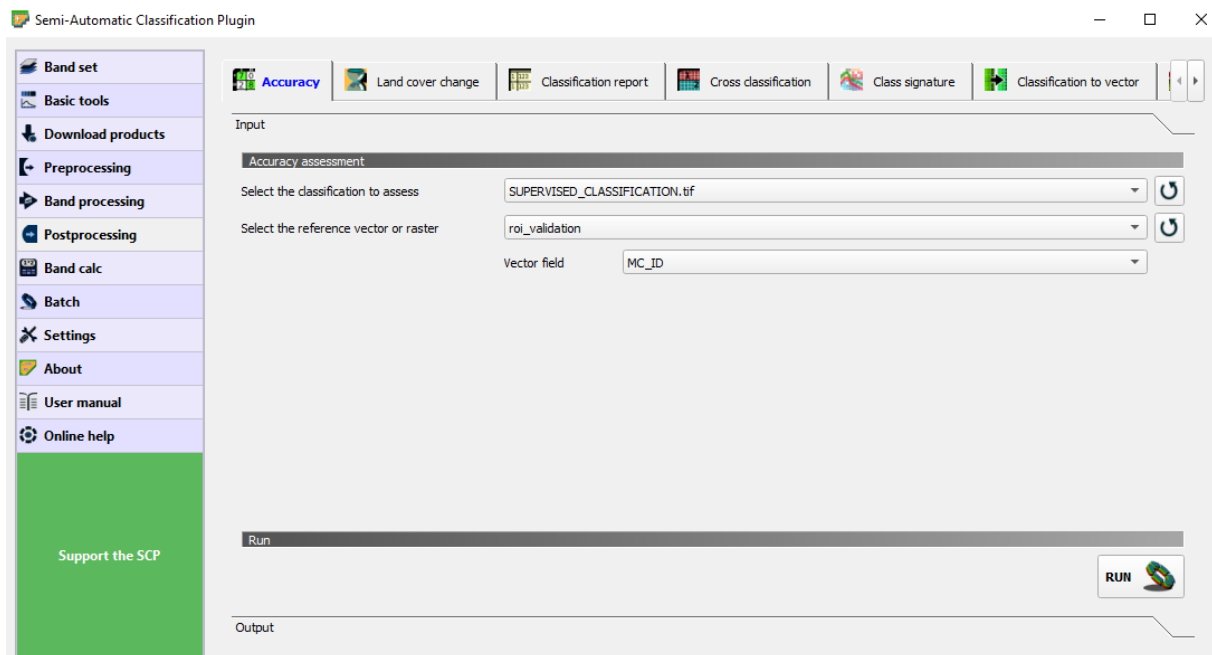
Figure 15. Accuracy assessment tool.

The output of the *Accuracy Assessment* tool is an Error matrix, where you receive information about the correctly classified pixel into appropriate classes. Several metrics may be calculated from the matrix. The basic one is Overall Accuracy together with two different error types. Those errors may be represented via Commission and Omission or Producer's and User's accuracy, Kappa coefficient may be also calculated complementary (Figure 16).

```
5                    0              0              0              0              679            679
Total                2130           311            648            775            1168           5032

                     > AREA BASED ERROR MATRIX
                     > Reference
V_Classified         1              2              3              4              5              Area
1                    0.1125         0.0000         0.0000         0.0000         0.0000         1356886700.0000
2                    0.0000         0.0240         0.0000         0.0026         0.0000         319912000.0000
3                    0.0000         0.0000         0.1635         0.0070         0.0000         2055479800.0000
4                    0.0283         0.1133         0.0185         0.3193         0.2099         8309253800.0000
5                    0.0000         0.0000         0.0000         0.0000         0.0012         14507700.0000
Total                0.1409         0.1373         0.1819         0.3289         0.2111         12056040000.0000
Area                 1698362884     1655055965     2193261819     3964823545     2544535787     12056040000
SE                   0.0034         0.0065         0.0031         0.0088         0.0079
SE area              41172884       77996202       37227582       105526108      95446494
95% CI area          80698853       152872555      72966060       206831172      187075129
PA  [%]              79.8938        17.4708        89.8563        97.0880        0.5702
UA  [%]              100.0000       90.3846        95.8796        46.3263        100.0000
Kappa hat            1.0000         0.8885         0.9496         0.2003         1.0000

Overall accuracy [%] = 62.0494
Kappa hat classification = 0.4748

Area unit = metre^2
SE = standard error
CI = confidence interval
PA = producer's accuracy
UA = user's accuracy
```
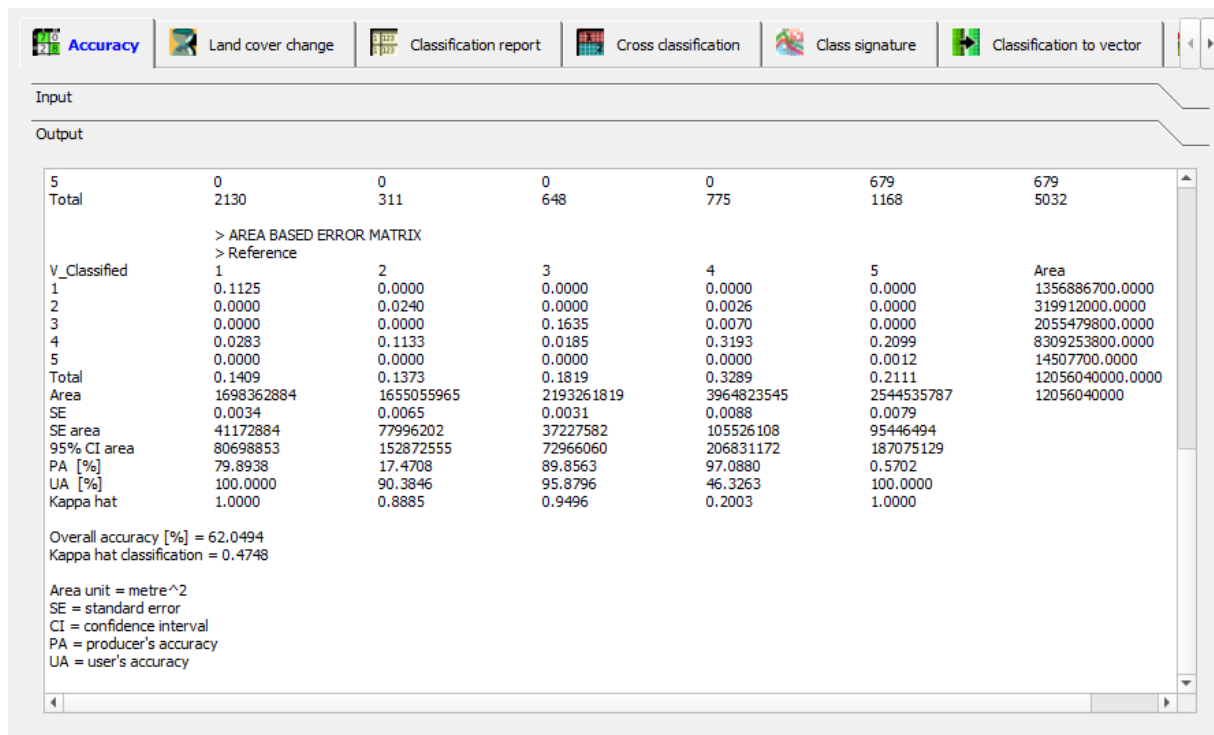
Figure 16. Error matrix.

As could be read from the matrix above, the rows stand for classified data and columns are representing reference data. The overall accuracy is quite high, i.e. above 60 %, however, some of other metrics seem pretty low, like PA of 17 %. Let's explain what each metric means. Find below (Table 1) the identical Error matrix as above, however, a little bit simplified.

Table 1. Simplified error matrix.

Note: Number of reference and classified data are expressed in a percentage of area, i.e. for example 0.113 means that 11.3 % of class 1 area was correctly classified as class 1. The area could be replaced by the number of pixels.

Overall accuracy (OA) is nice metrics, easy to calculate and understand but ultimately only provides the map user and producer with basic accuracy information. Overall accuracy tells us out of all of the reference data what proportion were mapped correctly. It used to be expressed as a percentage, where 100 % means perfect classification where all reference data was classified correctly.

In our case, OA is calculated like: (0.1125 + 0.024 + 0.1635 + 0.3193 + 0.0012)/1, i.e. 62.04 %.

Producer's accuracy (PA) is calculated like the number of reference sites classified accurately divided by the total number of reference sites for that class. This refers to the accuracy from the point of view of the producer. In other words, PA represents how often are real features on the ground correctly shown on the classified map or the probability that a certain land cover of an area on the ground is classified as such. The Producer's accuracy complements to the Error of the Omission, it means that 100 % - Omission Error = PA.

In our case, PA for class 1 is calculated like: 0.1125/0.1409, i.e. 79.84 %.

User's accuracy (UA) is calculated like the total number of correct classifications for a particular class and divided it by the row total This refers to the accuracy from the point of view of a map user. In other words, UA represents how often the class on the map will be present on the ground, this is referred to as reliability. The User's accuracy complements to the Error of the Commission, it means that 100 % - Commission Error = UA.

In our case, UA for class 1 is calculated like: 0.1125/0.1125, i.e. 100.00 %.

Kappa value evaluates how well the classification performed as compared to randomly assigning values, i.e. if the classification is better than random. The Kappa can range from -1 to 1, where a value of 0 indicated that the classification is no better than a random classification, a negative number indicates the classification is even worse than random. A value close to 1 indicates that the classification is significantly better than random.

The user's and producer's accuracies for any given class typically are not the same. In our case, the PA for class 1 is 80 % while the user's accuracy is 100 %. This means that 80 % of the reference class 1 areas have been correctly identified as class 1, but 100 % per cent of the areas identified as class 1 in the classification were class 1. However, if you take a look e.g. at class 4, PA is 97 %, while UA is only 46 %. It means that class 1 (0.0283), class 2 (0.1133), class 3 (0.0185) and class 5 (0.2099) areas were mistakenly classified as class 4. **By analysing the various accuracy and error metrics we can better evaluate the analysis and classification results**. Often you might have very high accuracy for certain classes, while others may have poor accuracy. The information is important so you and other users can evaluate how appropriate it is to use the classified map.