GEO TUTORIAL

#QGIS
Dealing with Coastal Flooding, part 3B:
USING SUPERVISED MACHINE LEARNING FOR LAND USE LAND COVER CLASSIFICATION

Krzysztof Raczynski
Flavia Xavier
John Cartwright

Geosystems Research Institute
Mississippi State University

MAY 2025

This work was supported through funding by the National Oceanic and Atmospheric Administration Regional Geospatial Modeling Grant, Award # NA19NOS4730207.

Geosystems Research Institute, Mississippi State University, Box 9627, Mississippi State, MS 39762-9652

The Geospatial Education and Outreach Project (GEO Project) is a collaborative effort among the Geosystems Research Institute (GRI), the Northern Gulf Institute (a NOAA Cooperative Institute), and the Mississippi State University Extension Service. The purpose of the project is to serve as the primary source for geospatial education and technical information for Mississippi.

The GEO Project provides training and technical assistance in the use, application, and implementation of geographic information systems (GIS), remote sensing, and global positioning systems for the geospatial community of Mississippi. The purpose of the GEO Tutorial series is to support educational project activities and enhance geospatial workshops offered by the GEO Project. Each tutorial provides practical solutions and instructions to solve a particular GIS challenge.

USING SUPERVISED MACHINE LEARNING FOR LAND USE LAND COVER CLASSIFICATION

Krzysztof Raczynski [1, 2, 4, 5, 6, 8] chrisr@gri.msstate.edu
Flavia Xavier [3, 7] flavia.xavier@msstate.edu
John Cartwright [7, 9, 10, 11] johnc@gri.msstate.edu

Geosystems Research Institute
Mississippi State University

CRediT: 1: Conceptualization; 2: Methodology; 3: Verification; 4: Resources; 5: Data Curation; 6: Writing - Original Draft; 7: Writing - Review; 8: Visualization; 9: Supervision; 10: Project administration; 11: Funding acquisition

REQUIRED RESOURCES

- QGIS 3+

FEATURED DATA SOURCES

- Click here to access dataset used in this tutorial (2.973 MB).

OVERVIEW

Coastal areas across the United States face increasing challenges from changing water levels, which can lead to more frequent flooding and infrastructure strain. In communities like Bay St. Louis, Mississippi, rising water can make roads impassable, damage property, and disrupt daily life—posing serious concerns for homeowners and local economies.

As part of a planning team, your role is to assess how changing sea levels may impact the safety, infrastructure, and long-term growth of this Gulf Coast community. The focus is on protecting property, ensuring economic stability, and strengthening community resilience. This is the theme of the *Dealing with Coastal Flooding* tutorial series, which includes the following topics:

- Part 1: Creating Raster DEM from LiDAR Data
- Part 2: Spatial Predicates: Preparing Residential Data
- Part 3A: Using Unsupervised Machine Learning for Land Use Land Cover Classification
- **Part 3B: Using Supervised Machine Learning for Land Use Land Cover Classification**
- Part 4: Hydrologic Raster Preparation: Resampling and Burning Stream Network
- Part 5: Generating Flooding Extent with Raster Calculator
- Part 6: Calculating Spatial Statistics of Inundated Areas
- Part 7: Creating 3D Maps of Flooding Projections
- Part 8: 3D Map Animations
- Part 9: Creating and Animating Timeseries

In the previous part, we processed spatial information to prepare residential data for further analysis. In this part, we will use machine learning to classify Land Use Land Cover (LULC) within our studied region. To determine different LULC classes, we will need a satellite image with different bands information, representing varying wavelengths of light captured by the remote sensors. There are two approaches we can take here to determine LULC classes automatically, that is, supervised and unsupervised machine learning. In the former case, we need to provide the algorithm with a training set, where we will tell it which areas are of which predetermined classes. The algorithm will use this information to extrapolate the dependencies over the area and classify each pixel according to similarities to the data we have prepared. In the second case, we do not predefine classes but let the algorithm decide which pixels are similar to each other. We will later decide what each detected class represents. In general, a supervised approach tends to provide more accurate results, while an unsupervised approach is useful when the final categories are unknown or when we want to perform a quick categorization. In this tutorial, we will apply the supervised approach. Make sure to check the remaining tutorials in the series to learn more about the entire analysis process.

DATA

For this tutorial, we will use *Landsat* satellite imagery that you can download from USGS EarthExplorer (requires free account) or the **Featured Data Sources** link above. If you chose the former approach, set the *spatial extension* to the area of **Bay St. Louis, MS** and the *cloud cover* limit to ***0%*** to grab only a clear image. For the datasets, use Landsat Level 1. You can choose any of the dates, but it is recommended to use some for the ***spring*** or ***summer*** time to obtain better differences in vegetation, as some of the winter images may lead to lower precisions of the result. You should download either the entire dataset (which might be around 1GB) or at least **bands 1 to 7**. The data provided in the *Featured Data Sources* is .zip archived (remember to extract before use) and cropped to the 3-kilometer spatial extent of the building boundary data (upper left coordinate: 789990.8953, 316815.7082; lower right coordinate: 839412.6408, 266091.7101 in EPSG:6507) (Fig. 1). Once you have downloaded the necessary data from either source add all the seven bands (in .tif files) into your project in QGIS.

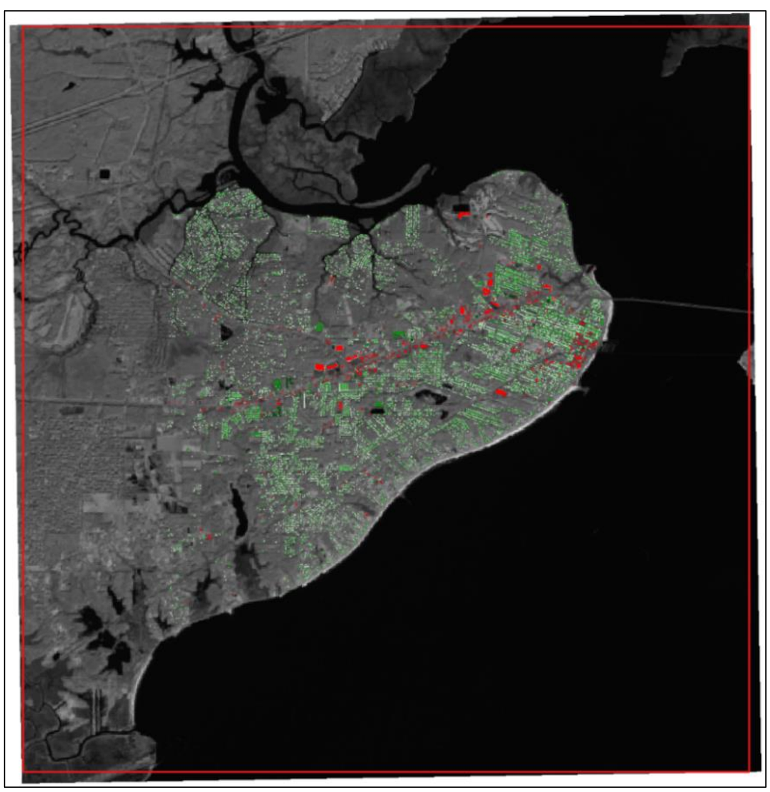

Fig. 1. *Landsat band 5 image against the 3-kilometer buffered building bounding box (in red) and the study area buildings generated in previous part.*

PLUGIN INSTALLATION

We can use *SAGA Tools* to run supervised classification if we have a layer presenting predefined classes. Alternatively, we can use an external plugin named *Semi-Automatic Classification* to perform more interactive supervised machine learning.

To install the plugin, open the *Plugins* menu and select *Manage and Install Plugins*. In the *All* tab, type *classification* in the search box, locate the plugin named ***Semi-Automatic Classification Plugin (SCP)***, authored by *Luca Congedo*, and proceed with installation.

This plugin offers a wide range of useful tools and scripts, so be sure to test it beyond the scope of this tutorial. Once the plugin is installed, you should see the *SCP* menu in the menu bar and the *SCP Dock* panel attached somewhere in your main window (Fig. 2). If you do, you can proceed to the next section.

It might happen, however, that instead of the plugin panels you will see an orange warning about *Remotior-Sensus* (or other dependency either not found or outdated). If this is the case, you need to install it manually. In your *Windows* search bar or program catalog (for systems different than *Windows*), search and open *OSGeo4W Shell* (there is a good chance it will be in your QGIS app folder). After opening the shell type:

```
pip3 install --upgrade remotior-sensus
```

If your QGIS error mentioned more missing dependencies, you can install them all together e.g., to install *remotior-sensus*, *scikit-learn* and *torch* altogether type:

```
pip3 install --upgrade remotior-sensus scikit-learn torch
```

After installation is complete, restart QGIS.

IMAGE CONVERSION

When using satellite imagery, like *Landsat* images, for classification of land use and land cover (LULC), we first need to convert the *raw digital numbers* (DNs) to *Top of Atmosphere* (TOA) *reflectance*. This conversion ensures that the data represent consistent and comparable surface conditions. Raw *Landsat* data consists of pixel values that depend on sensor-specific settings and conditions at the time of image acquisition, such as the sun's angle, Earth-Sun distance, and atmospheric effects. By converting to *TOA reflectance*, we adjust these pixel values to represent the proportion of solar radiation reflected by the Earth's surface as observed from space. This step ensures the data is consistent and comparable across different scenes, sensors, and acquisition dates, enabling more accurate classification of land use and land cover types while minimizing the influence of external factors.

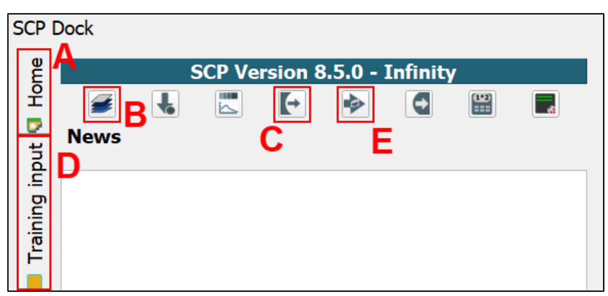

Fig. 2. SCP Dock main window.

To convert the raw images to *TOA reflectance*, we will use one of the *SCP* plugin options. First, locate the *SCP Dock* panel. If not visible, enable it from the *View* menu, under *Panels*. You can then drag and drop the panel like any other QGIS component to position it in the most desired location, e.g., dock it as a secondary tab in the *Processing Toolbox* area for easy access.

In the *SCP Dock* click on *Home* tab (Fig. 2–A) then select *Band set* (Fig. 2–B). A new window will open with the *Band set* tab (Fig. 3–A). If you did not load the input layers to QGIS, you can use the *Open a file* button (Fig. 3–B) to load files from the drive. If you have loaded the layers into QGIS, you can use the *Add bands loaded in QGIS* (Fig. 3–C). The middle window will list all seven band files once you add them. Make sure that they are in order from band 1 (top of the list) to band 7 (bottom of the list). Navigate to the bottom, where the *Band quick settings* are located, and change the *Wavelength* preset (Fig.

3-D) to *Landsat 8 OLI [bands 1, 2, 3, 4, 5, 6, 7]*. This will change the center wavelength settings. Check the *Create virtual raster of band set* (Fig. 3-E) and click *Run*. A *Select a directory* window will pop up. Choose a directory to host your data. In the *Semi-Automatic Classification Plugin* window, switch tabs on the left to *Preprocessing*. From the submenu, select *Image conversion*.

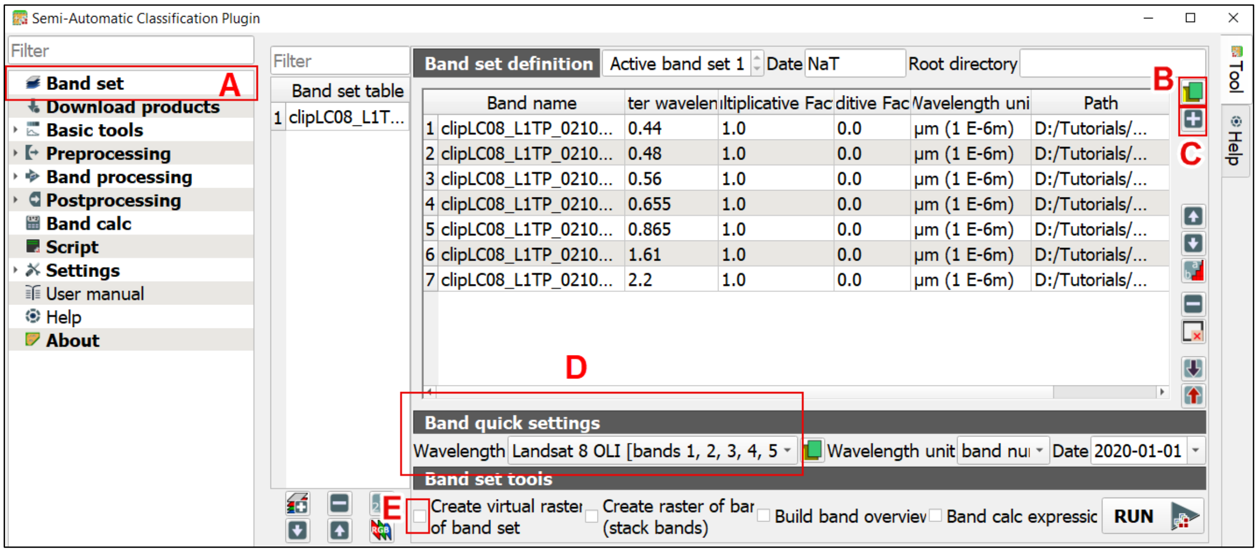

Fig. 3. *By default, class 2 in LiDAR data contains ground surface.*

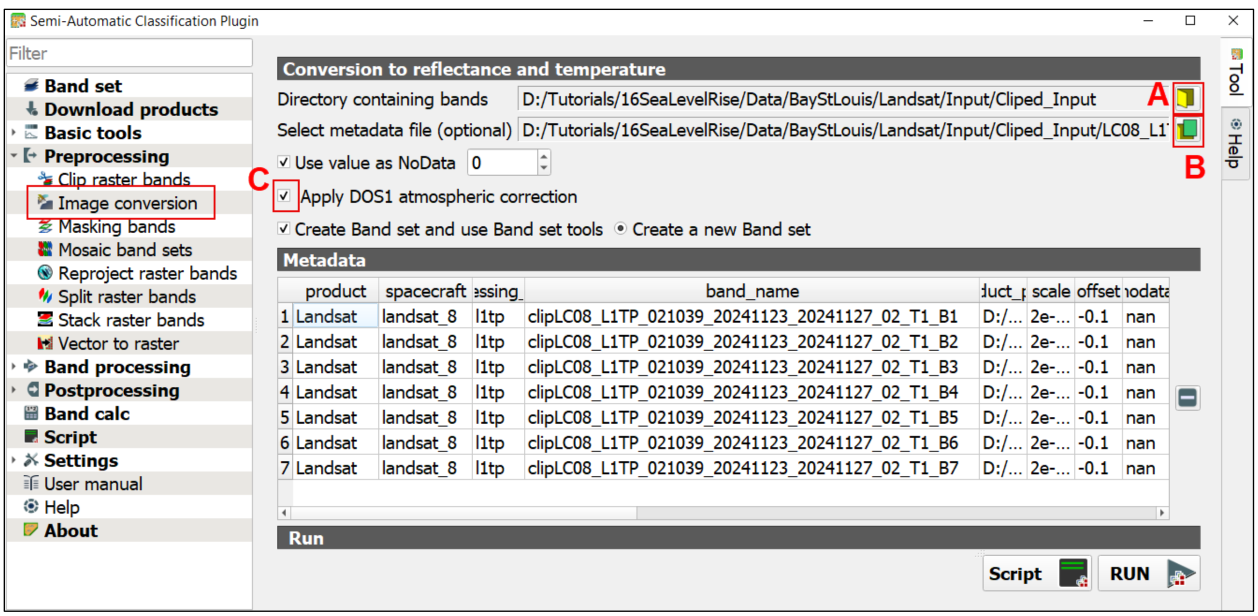

Fig. 4. *We need to convert the raw digital numbers to Top of Atmosphere reflectance to ensure consistent surface conditions.*

TOA CONVERSION

In the *Image conversion* tab, we will handle *conversions to reflectance and temperature*.

A. First, under *Directory containing bands*, click the *Select a directory button* (Fig. 4–A) and select the folder where your bands are.

B. Then, under *Select metadata file*, click the *Open a file* button (Fig. 4–B) and choose the ***MTL metadata file*** available with the tutorial data or your Landsat imagery. The list will load the band files once you select the metadata file. If the directory contains more raster files, the system will also load them. In such a case, select them and use the *Delete row* button to remove them from the list.

C. Check the ***Apply DOS1 Atmosphere Correction*** checkbox (Fig. 4–C) and ensure all bands 1–7 are on the list.

D. Then click *Run*. Note: the *Select a directory* window may pop up again so you can select the destination of the new files after processing.

After a moment, the process will be finalized. If your newly processed files are not automatically added, navigate to the folder where you saved the result and add them to the project.

E. To better visualize the area, add the *virtual raster* (*virtual_raster.vrt*) from the previous step. The virtual raster will be displayed in the scales of blue, because the color bands are not set properly.

F. Right-click the virtual raster layer and select *Properties*.

G. In the *Layer Properties* window, click the Symbology tab.

H. In the *Band Rendering* settings, set the *render type* to *Multiband color*. Change the *red band* to ***Band 4***, the *green band* to ***Band 3***, and the *blue band* to ***Band 2,*** and click OK.

I. Now go back to the *Band set* in SCP Dock (first icon in the Home tab), and verify that your *Band set* table displays ***RT_....*** rasters. If you have more than one set (true if you are using own data), you can select and remove it using the *Remove selected band set* button in the middle panel. You can close the *Band set* (*Semi-Automatic Classification Plugin*) window.

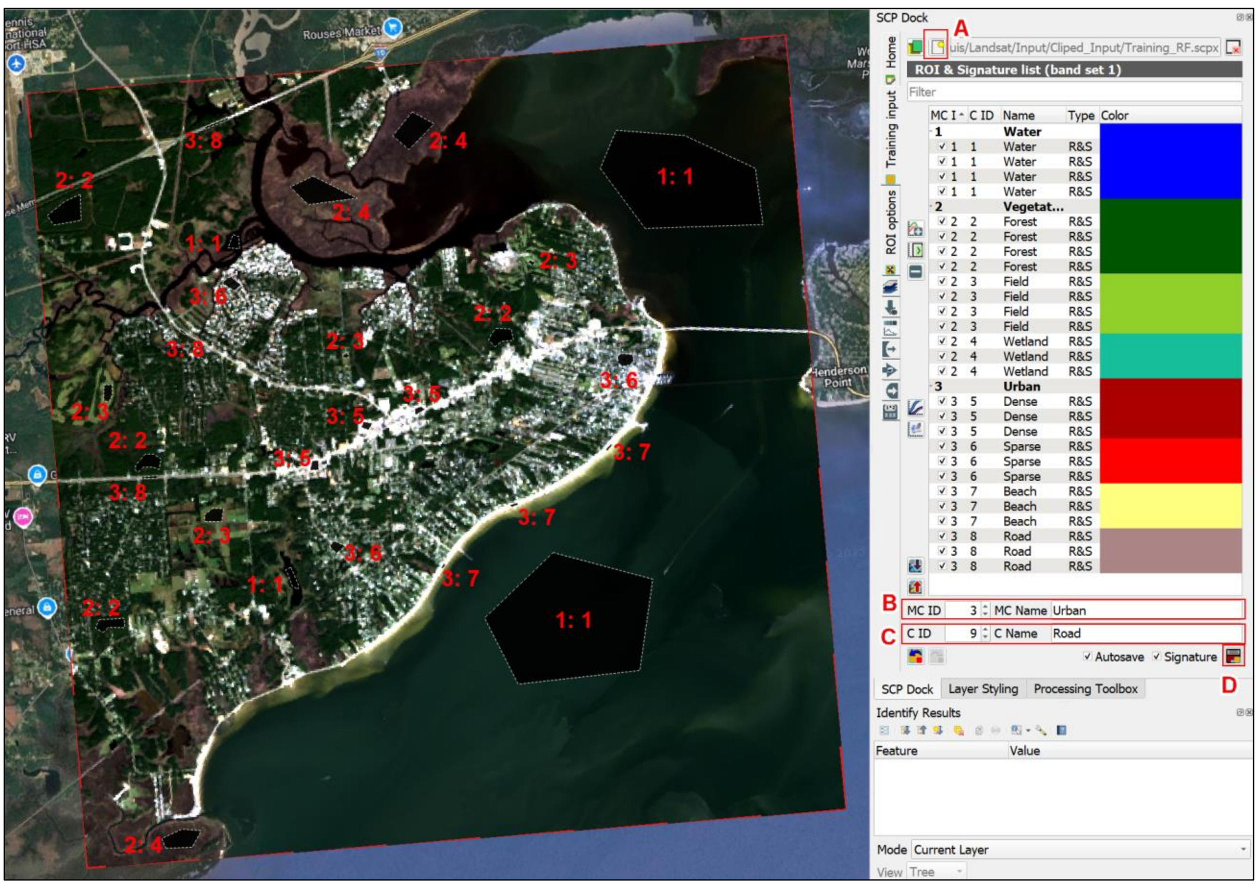

Fig. 5. *Example setup of training set for SCP training layer features locations with the assigned MC and C IDs as well as classification and colors with IDs used for each class. To view a larger version with higher contrast, click here.*

CREATING TRAINING SET

To use supervised classification, we need to prepare a training set that will indicate to our algorithm which regions represent a specific class of LULC. To do so:

A. On the SCP Pannel, switch the *SCP tab* to *Training Input* (Fig. 2–D).
B. Click *Create a new training input* (Fig. 5–A) and select a save location for the file. Now that the file is created, you should see the *SCP training layer* in the *Layers* panel.

Now we need to prepare training data, where we will draw polygons on the map and set the categories used for classification. Based on the LULC characteristics of our study area, we will classify the data into eight classes: *water*, *forest*, *fields* (or *grass*), *wetland*, *dense urban*, *sparse urban*, *beach, and road.* These classes will be grouped into three macroclasses: *water*, *vegetation*, *urban*. We need to decide the sampling locations for each class, based on our input data (Fig. 5). To do so, analyze the map and select areas that are representative for each class, and well visible in the imagery.

The training dataset preparation workflow is as follows:

A. In the *Layers panel*, select the *SCP training layer.*
B. Then, in the plugin toolbar, click on the *Create a ROI polygon* icon. Your mouse pointer will change to a cross indicating that you can start drawing a region of interest (ROI) on the map.
C. Choose the area that will be used as input and draw a polygon on the map (Fig. 5). Use the virtual raster to help in the drawing process. Tip: right-click after placing the final vertex to close your polygon.
D. In *Macroclass ID* (*MC ID*, Fig. 5–B), set a name for the macroclass e.g., 1-***Water***, 2-***Vegetation***, 3-***Urban***. Each time you use a new macroclass, change to the new *MC ID* number and name.
E. Set the *Class ID* (*C ID*, Fig. 5–C) accordingly to the LULC you want to use (the 8 classes named before).
F. After drawing a polygon and assigning the IDs, click *Save temporary ROI to training input* to store the sample (Fig. 5–D).
G. Once you have provided one polygon along with its *IDs* and *class names*, you can draw another polygon of the same class by switching the *C ID* back to the one you previously provided. If you need to delete a sample, select it in the SCP Dock (it you turn blue when highlighted), right-click on it, and choose *delete items*.
H. Use multiple polygons for each class to improve training data.

An example of a training dataset setup is presented in Fig. 6, with locations shown in Fig. 5. Remember, the more precise the training input, the better the output.

Once you have three to four polygons for each class, set its color. You can select multiple rows and then right-click to invoke the context menu. From there, you can use the *Change Color* option to adjust the class colors for the output raster (Fig. 5).

CLASSIFICATION

Switch back the *SCP* tab to *Home* (Fig. 2–A) and select *Band processing* (Fig. 2–E). In the submenu, select *Classification* (Fig. 6–A). For the *input* settings, check the ***use input normalization*** box, and select ***Class ID*** for *Use training*. Normalization usually improves the machine learning model's capabilities, as all input variables are within a common range (typically 0 to 1) instead of having varying ranges. There are multiple algorithms available, and we encourage you to test different settings. For this first run we will use ***Random Forest***; click its tab. You can adjust various additional settings for the algorithm, but a key consideration is the balance between model parameters for accuracy and computational costs. When it comes to *Random Forest*, there's no universal guideline for the *number of trees* and *splits* to utilize. However, numerous studies indicate that boosting the initial number typically enhances the output, but only to a certain extent. Beyond that, the improvement is minimal compared to a lower setting, but the computational requirements escalate. You are welcome to modify any parameters to see how they affect the outputs. In this case, we will use ***16 trees*** with a minimum number of ***4 splits***. Additionally, use the ***cross-validation*** setting and ***calculate the classification confidence raster*** (Fig. 6). Once all the settings are in place, click *Run*. A *Save classification* window will pop up. Select your preferred directory to save your output image and define its name. Computation time will heavily depend on your computer processing capabilities and parameters used, but once the process is finished, you will see a classification image in the map (Fig. 7).

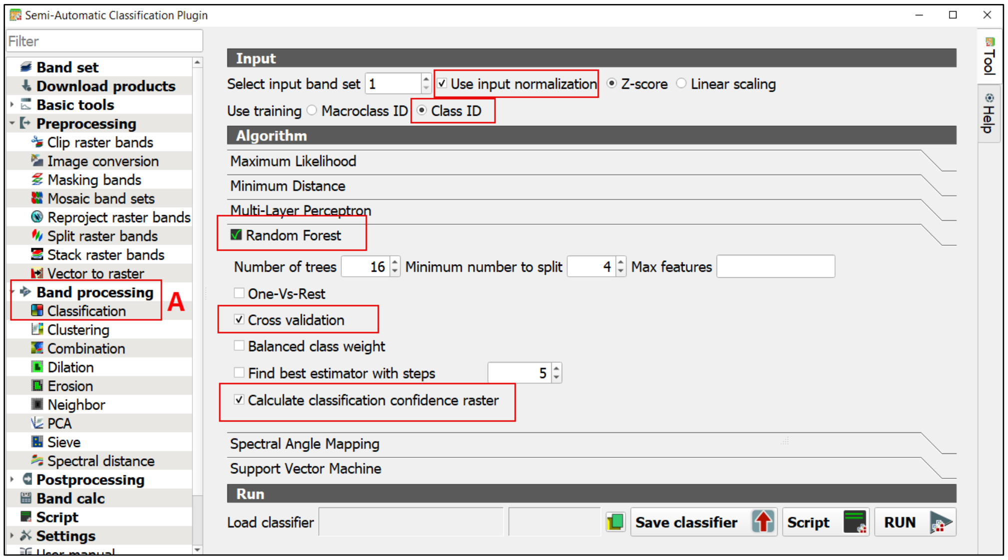

Fig. 6. Example of Random Forest classification settings.

The accuracy of the output will depend mainly on the algorithm and parameters used, as well as the quality of the training dataset prepared in the previous step. You can check how well the algorithm performed and the regions where classification struggled most by displaying the calculated *confidence raster* (Fig. 8). This image presents the confidence with which a classifying algorithm assigned the specific class to each pixel. Values of 1 indicate 100% confidence, while the lower the value, the lower the confidence of a given region being classified to a given class. A quick analysis of the confidence levels and spatial extent indicates that the algorithm struggled most for sparse urban areas. This is expected due to a high variability between pixels representing roofs and roads classified as urban, and trees and grass classified as vegetation.

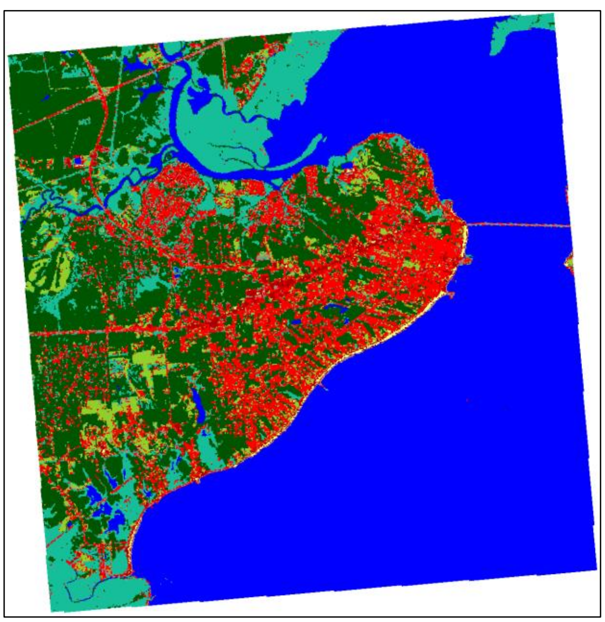

Fig. 7. The resulting raster with predefined classes detected by Random Forest algorithm.

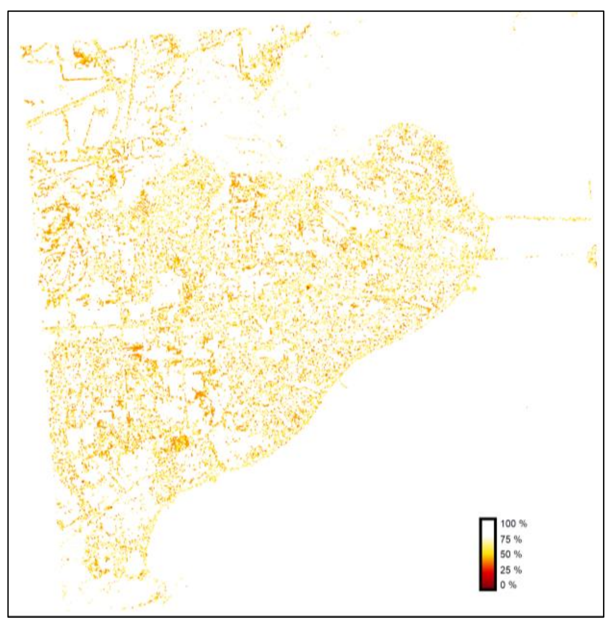

Fig. 8. Classification confidence presenting areas, where algorithm struggled most with classification.

EXERCISE

Analyze where the algorithm had the lowest confidence for classifying LULC and improve output quality by providing a more accurate training dataset. Try different algorithms to test their performance.

CONCLUSION

Congratulations on completing this GEO Tutorial in which you applied supervised machine learning techniques to classify LULC using Landsat satellite imagery! By converting raw data to TOA reflectance and leveraging the SCP in QGIS, you ensured accuracy and consistency in the results. This classification process is essential for understanding how different land uses may be impacted by sea level changes and for supporting effective flood risk analysis. Using this approach contributes to more informed decision making in coastal planning and resilience strategies. Be sure to explore the remaining tutorials and continue building your skills in coastal flood impact analysis!