Image Processing Lab Exercise #8:

Dr Bagher Bayat, RS/GIS Group, Faculty of Earth Sciences, SBU

Classification III: Supervised Classification, and Accuracy Assessment

Objectives:

- Utilize pre-existing (either from your fieldwork or Google Earth Pro) training sample data to create ROI's (regions of interest), which are regions in n-dimensional spectral space, also called feature space, defined by the centroids and variances/covariances of the values of the training sample data for each pre-determined class.
- Perform a supervised classification
- Conduct accuracy assessment for your classification
- Create manually your own ROIs directly from the image (when data from fieldwork or Google Earth Pro is not available)

The data for this lab include:

gnv-subset-etm-11-feb-2003-17-39.img – A Landsat 7 ETM+ image of Gainesville, FL.
2019_All_Training_Samples_All_Time_Crosswalk_in_situ.csv – over 1600 Training samples collected over multiple years for Gainesville.
Error_Analysis_Samples.csv – 258 Error Analysis samples developed by classes in the past.

NOTE: you may use a subset of the large TS file for error analysis to learn how to hold out a fraction, say ¼ or 1/3 of all of the data. The method is described in the instructions.

Introduction

Supervised classifiers require input of class information before classification, e.g. the training sample points we collected. There are multiple types of supervised classification methods, we will focus only on a couple of the types. Below all the types are discussed including: Parallelepiped, Minimum Distance, Maximum Likelihood, and Mahalanobis Distance (note that there are more methods are available such as *Binary Encoding, Neural network, Decision Tree, etc*).

With the *Parallelepiped* classifier, ENVI constructs an n-dimensional parallelogram, called a parallelepiped, in n-D spectral space (where *n* represents the number of bands being used in the classification), wholly enclosing each ROI. ENVI assigns only those pixels that fall into the parallelepiped into the class corresponding to the ROI. By leaving the default parameters in place, this method frequently leaves many pixels unclassified.

The *Minimum Distance* classifier calculates the distance of each pixel in n-D spectral space from the n-D means of the ROIs. The pixel is assigned to the class corresponding to the ROI with the closest mean. If the pixel is beyond a threshold distance, if you have assigned one, from the closest mean, it remains unclassified.

The *Maximum Likelihood* classifier uses the distribution of data within each ROI to calculate n-D probability functions for each class. Each pixel is assigned to the class for which the highest probability is calculated. This classification works best when the data are normally distributed. Again, it is common to have many pixels remain unclassified with this method.

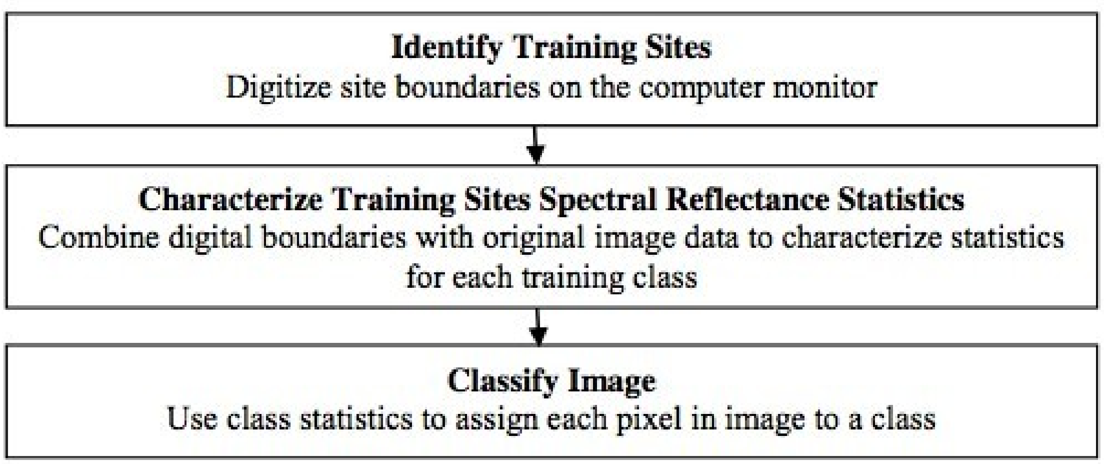
The *Mahalanobis Distance* classifier is a direction-sensitive distance classifier. It is similar to Maximum Likelihood, except that it assumes all class covariances are equal and is therefore a faster method.

When entering classification parameters for supervised classifiers, you have the option to designate a single threshold for all classes, a separate threshold for each class, or no threshold. For example, when performing a Parallelepiped classification, you can specify different widths in each band for each class based on reflectance standard deviation. This allows more control over classes where you have confidence that the training data are more clearly defined and those that are less distinct.

Post classification is where the true analysis and interpretation begins. In this lab we will conduct accuracy assessments. We have started to see some of this in the unsupervised classification when we performed the class statistics function. These analyses will become more in depth with this lab.

Part I: Supervised Classification

To perform a supervised classification, first we identify small areas for which we know the land cover. These areas are called training sites. The image brightness values within the training sites are generalized (n-dimensional centroids and variance/covariances are calculated) to represent the typical spectral properties of the training classes. Finally, the rest of the image is compared on a pixel basis to the generalized training-class data, to determine to which class each pixel belongs. The figure below illustrates the steps involved in supervised classification.



We have already obtained a file with training sample data and played with this in the previous lab (Lab 7). So, our first step is to take that data from single points of land cover classes to polygons. The process of converting from points to polygons, which we have made by hand, is called digitization. Before we begin digitizing, a general background and tips regarding digitizing may help improve the effectiveness of your work.

- Select typical areas. The training polygons are selected to represent the typical properties of the class. Thus it is important that your training sites should be as homogenous as possible. Meaning that your polygons should contain only one land cover type, no mixels! [*mixed pixels* ☺]. You will be working with the pre-existing data collected in multiple previous years (your csv file from last week's lab). Also, remember when you plan using this CIPEC TS form you can use 90 x 90 m continuous areas therefore this should be the smallest digitized polygon although the polygon should be smaller if your area is smaller than 90 x 90. Conversely, do not make the area too large or the problem of spatial autocorrelation will become important. We recommend areas between 4 and 25 pixels. Know that this is arbitrary but it works for us.

- Digitize efficiently. Do not spend hours digitizing each class. It is important to be able to outline a typical area rapidly. You should zoom into a specific area, this will make the digitizing easier. However, it is not worthwhile to zoom in so much that you are digitizing individual pixels.

- Digitize a reasonable number of pixels. Your training samples should contain an adequate sample of pixels for statistical characterization. Don't do too many, though, because adjacent pixels are always spatially autocorrelated.

- Bear in mind the difference between spectral classes and information classes. For example, assume you have two water bodies, one with clear and deep water, and one with shallow, muddy water. That might give rise to two very different spectral classes (Clear water, and Muddy water). We could regard them as one informational class (Water). Please note the difference between the two terms: spectral classes, and informational classes.

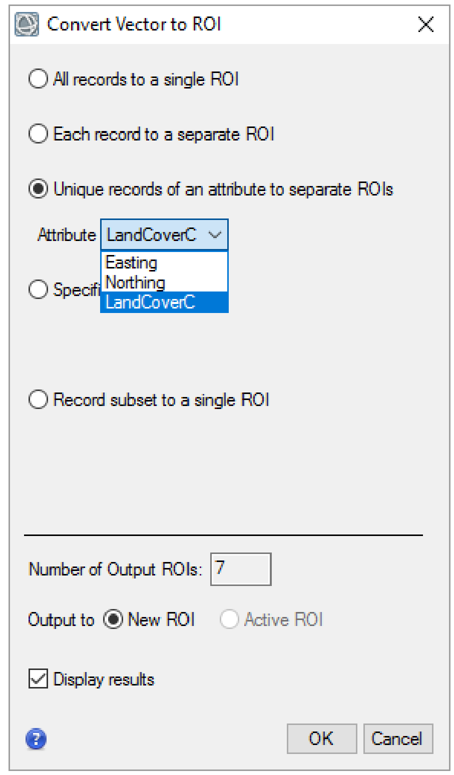
In ENVI we will digitize our points to spatial Regions of Interest (ROIs) or shapefiles. These points or polygons, which some software are called Areas of Interest (AOI) are the basis of the supervised classification. ROIs can be generated from numerous sources: manually from an image using the mouse, from scatter plots, and from vector layers, in this lab we will both manually create the polygons and use a pre-existing point file.

1. Open ENVI and add the Gainesville ETM image (5,4,3) and the shapefile of the projected training samples that we have provided named:
"All_Training_Samples_All_Time_Crosswalked_project.shp". To open the shapefile just drag and drop the .shp file into the ENVI interface.
Note that this shapefile is created from the original csv file named
"2019_All_Training_Samples_All_Time_Crosswalk_in_situ.csv" (and you know how to convert this csv to a point shapefile with ArcGIS from last lab)
You might have to increase the brightness of the image to allow full visualization. Make sure that in the display the training samples are visible on top of the image. The land-cover names are the consensus of many classes over many years:

a	Bare
b	Forest
c	Grassland
d	Pasture
e	Urban
f	Water
g	Wetland

You may use these data as you wish: as training samples, as error analysis samples, or both (although use different points for training and error analysis).

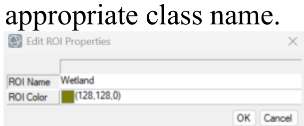
2. The All_Training_Samples_All_Time_Crosswalked_project.shp file has over 1600 points from the Gainesville area. We will use all of these points as training samples and then use the second shapefile *Error_Analysis_Samples.shp,* for the error analysis. These two shapefiles are provided in the input folder and we need to create a ROI file using ENVI Toolbox: Regions of Interest: **Vector to ROI**.

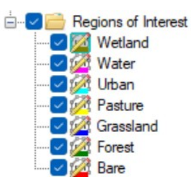


Warning here: be sure to use "Unique records of an attribute to separate ROIs." And "LandCoverC" in the Attribute.

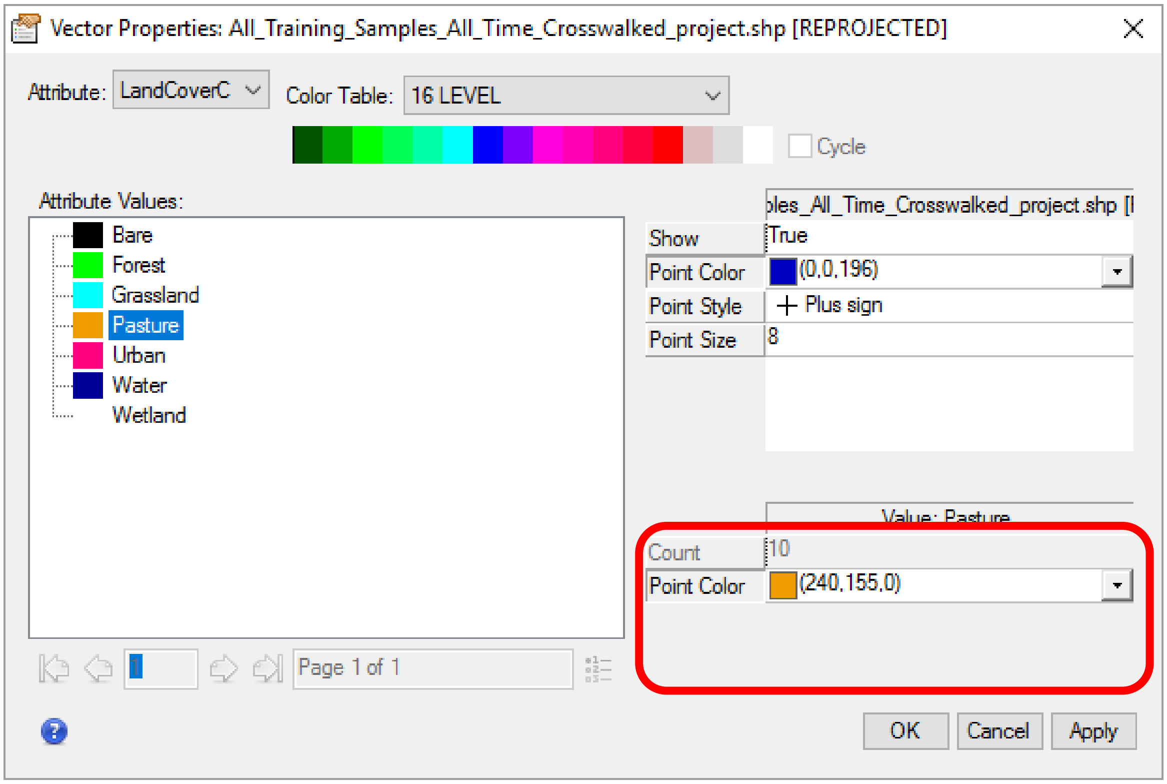
This creates separate ROI files for each land cover type, which is what you will need later. Each ROI will be comprised of all the pixels that contained the points in the shapefile.

After generating the ROIs, you'll need to rename them according to their corresponding class names in ENVI. To do this, right-click on the ROI layer name in the Layer Manager and select *Properties*. In the *Edit ROI Properties* window, update the name field to reflect the appropriate class name.





3.Good to know that, if you wish you can change the display symbology of our points. Right-click your training sample shapefile and select properties. In the properties dialog find the select attribute window, change this to LCCLASS (or LandCoverC). You can change the color of the points on the screen Using the Value: Point Color field (below). Optionally you can change the color table and the symbols as you see fit.



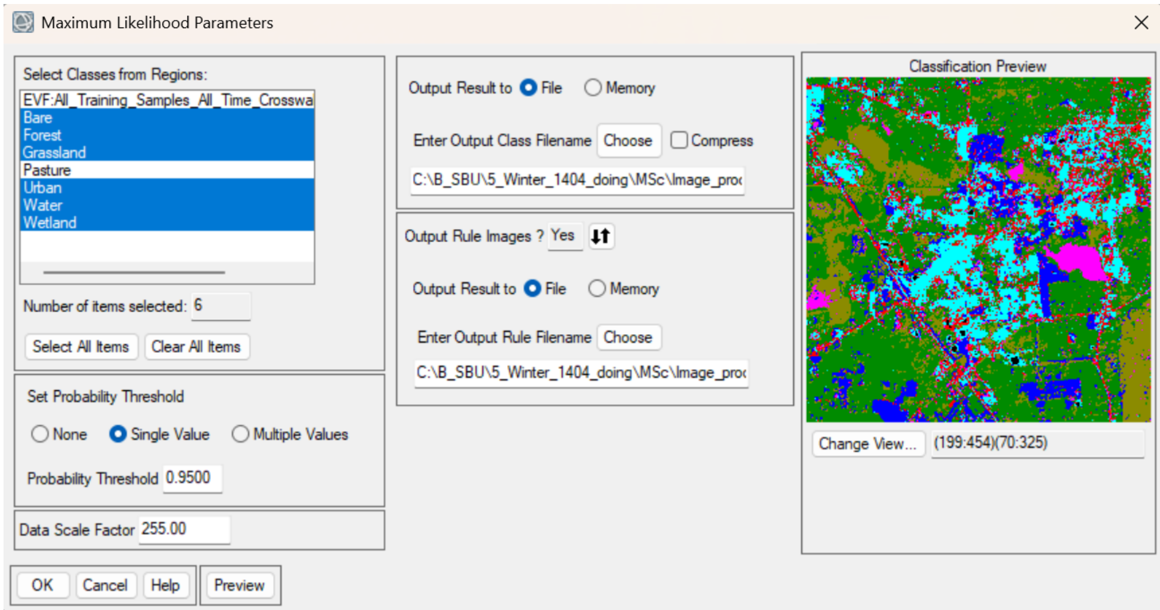
3. In the ENVI Toolbox expand the **Classification and Supervised Classification** folders. Then double click on **Maximum Likelihood Classification**. A Classification Input File dialog appears.

4. Select the Gainesville image as the input. Click **OK**. Note that we are leaving in the high- and low-gain thermal bands (6 and 7). These bands may help in your classification especially in residential areas. The thermal band is often omitted in classifications, even though it is highly useful.

5. The Maximum Likelihood Parameters dialog appears. The ROIs you created should be listed. Select the ROIs that you wish to use. Select all of the classes except Pasture (despite the very large number of TS, there are still not enough Pasture for the algorithm to work). You might want to separate out a single land cover for helping with cluster-busting an unsupervised classification, in which case you would choose a single category. To set a threshold so that some pixels are unclassified, enter a Probability Threshold value of 0.95. This means that if the highest likelihood calculated for all classes for a given pixel is less than 0.95, then the pixel will remain unclassified.

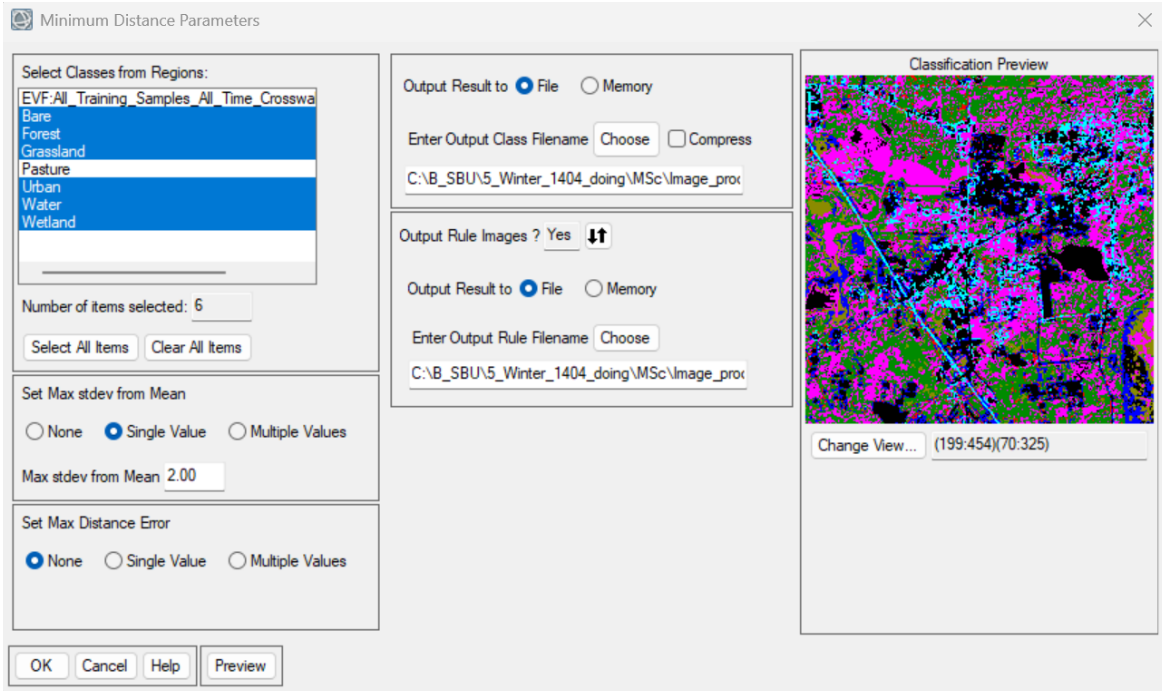
6. Click the **Preview button**. The image that appears helps evaluate the classes you have created. Does it look reasonable?

7. For Output Classification Filename type Gainesville_maxlike_supervised_11_Feb_2003.dat. For Output Rule Filename type Gainesville_maxlike_rule_11_Feb_2003.

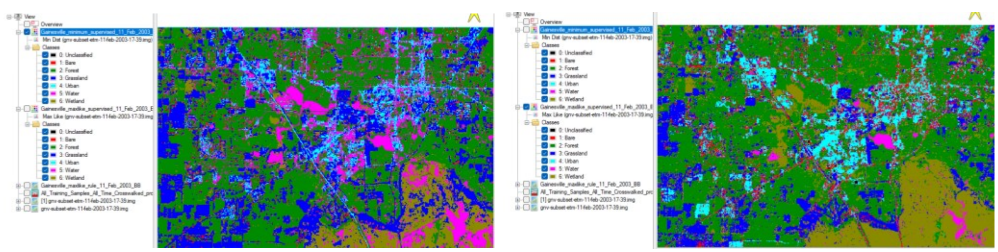


8. Click **OK** to run the classification. After a minute or so, the land-cover classification map will be displayed in the viewer.

Follow a similar method as above but this time run a supervised classification using a minimum distance algorithm. Once again, do not use Pasture. This algorithm could use it, but it would make comparisons more difficult. It is your decision whether to change the thresholds in supervised classification dialog. You can choose "None" for now. If "None" is selected, ENVI assigns every pixel to the spectrally nearest class, regardless of how large the spectral distance is. If a threshold is specified, pixels that are farther than the threshold from all class means remain unclassified. Alternatively, you can try playing with a few values (for example try Max Stdev from Mean: Single Value = 2 and Max Distance Error: None). Remember our goal is to produce the most accurate classification possible. Name the output Gainesville_minimum_supervised_11_Feb_2003.dat and Gainesville_minimum_rule_11_Feb_2003.



Turn off the Rule images in the Layer Manager, change the colors of the classes in each classification image to match classes with the same colors, and blink back and forth between the Maximum Likelihood and the Minimum Distance images. Compare the two images both visually and with the Cursor Value function.



Note: A Rule Image is an output that shows how similar each pixel is to each class. In the Minimum Distance classifier, this similarity is determined by the spectral distance between the pixel and the mean spectrum of each class. The smaller the distance, the greater the similarity of the pixel to that class. While the classified image shows only the final class assigned to each pixel, the Rule Image indicates how strongly or confidently each pixel matches every class.

In Maximum Likelihood classification, the Rule Image contains discriminant function (log-likelihood) values rather than probabilities. These values may be positive or negative. The important point is that the pixel is assigned to the class with the highest rule value. Therefore, negative values are perfectly normal; they simply indicate a lower likelihood relative to the class with the highest value

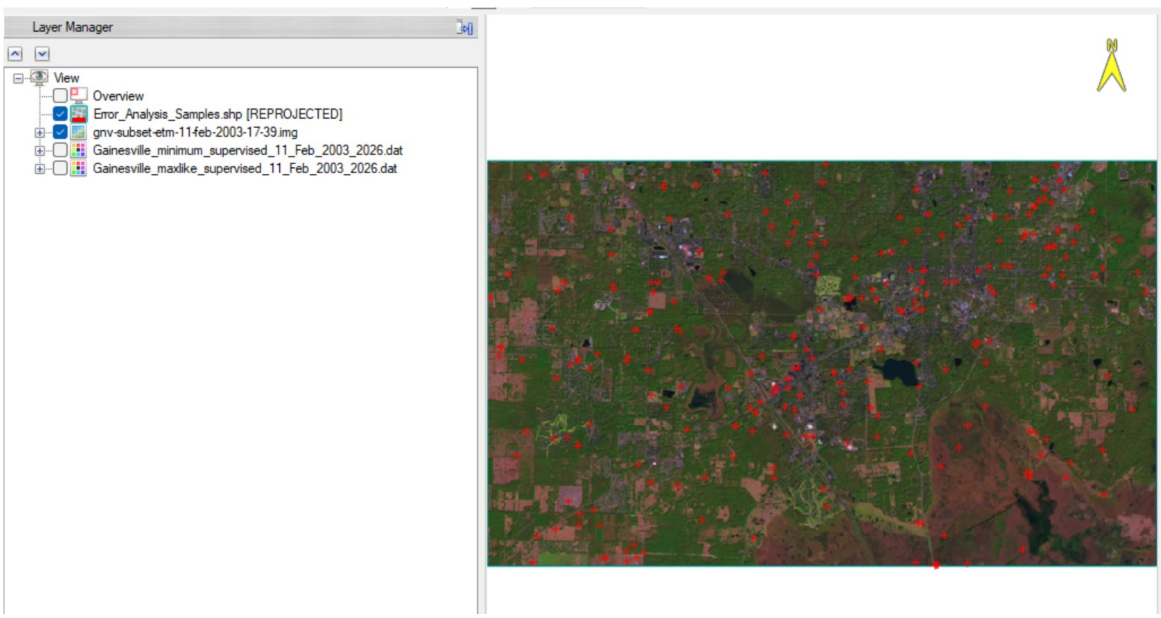
Part II: Accuracy Assessment

Confusion Matrix

A classification accuracy assessment requires two sources of information: pixels in a remote sensing-derived classification map, and ground reference test information. In this section, two files will be used: the first file is a ROI accuracy assessment file (*Error_Analysis_Samples.shp* or your subset from the larger collection of training samples) in your lab 8 folder, the other is your classification you produced above.

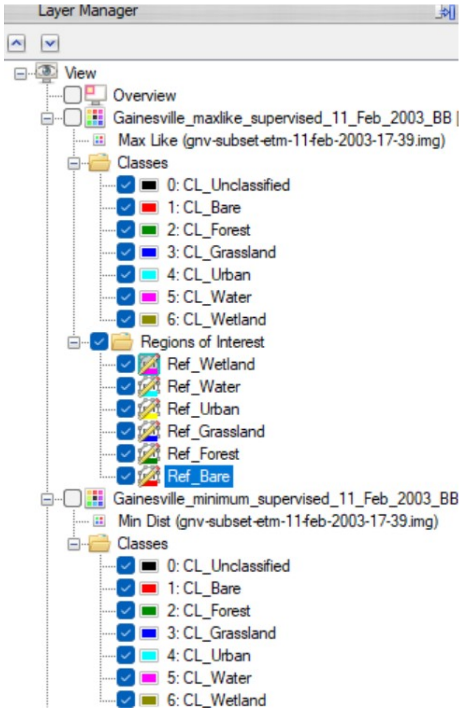
ENVI's confusion matrix function allows for comparison of two classified image or a classified image and ROIs. You will perform two different accuracy assessments, one for each type of supervised classification performed.

1. Keep the two classified maps and load the Gainesville ETM image (5,4,3) if it Is not already open and also load your accuracy assessment shapefile. Simply drag and drop the .shp file into the ENVI interface



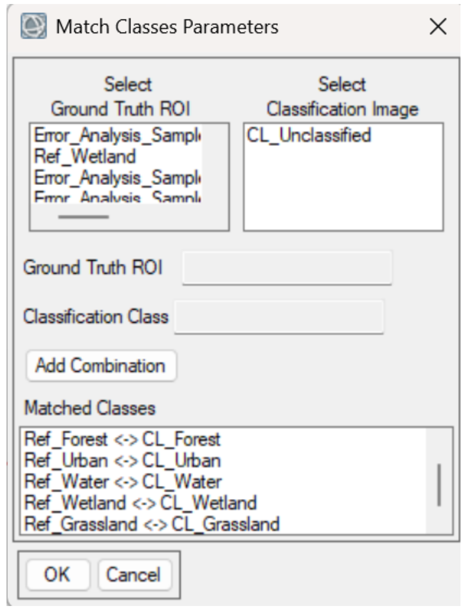
2. Convert the .shp file to an ROI file with ENVI Toolbox: Regions of Interest: Vector to ROI. The Regions of Interest file based on points will appear in the Layer Manager. Again specify Unique records of an attribute to separate ROIs and select the ETM image or maxlike classification as the visualization layer. Click OK.

3. Now to make things easier to match classes with their error analysis points, Edit the Class Names in Edit Envi Header to be CL_Bare, CL_Forest, etc., instead of their long names, for the Classification, and change the accuracy assessment ROI names to Ref_Bare, Ref_Forest, etc.



4. Select Classification > Post Classification > Confusion Matrix using Ground Truth ROIs Remember that "Ground Truth" is a misnomer. If you do not believe that, look closely at some of the points. There are error analysis points named "Forest" that appear in Grasslands. Why do you suppose that is? Even TS are measured with error and are therefore NOT "truth.")

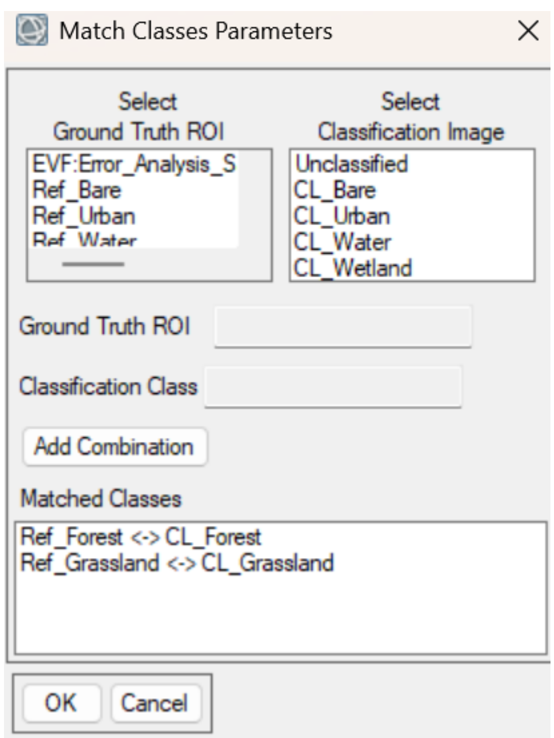
5. For the ROI method, select the classified image to be evaluated (if you do not already have it open select open in the confusion matrix dialog and browse to your lab 8 folder). Match the image classes to the ROIs previously loaded. This is a bit tricky because the window size cannot be increased so you must use the slider to look at the names of the classes.



When all the matches are complete, click OK. Choose both Pixels and Percent output and check Yes for Report Accuracy Assessment.

6. Examine the confusion matrix

7. Repeat step 5, but this time select only the Forest and Grassland classes to generate the confusion matrix and perform the accuracy assessment. Can you explain why the classification accuracy improved in this case?

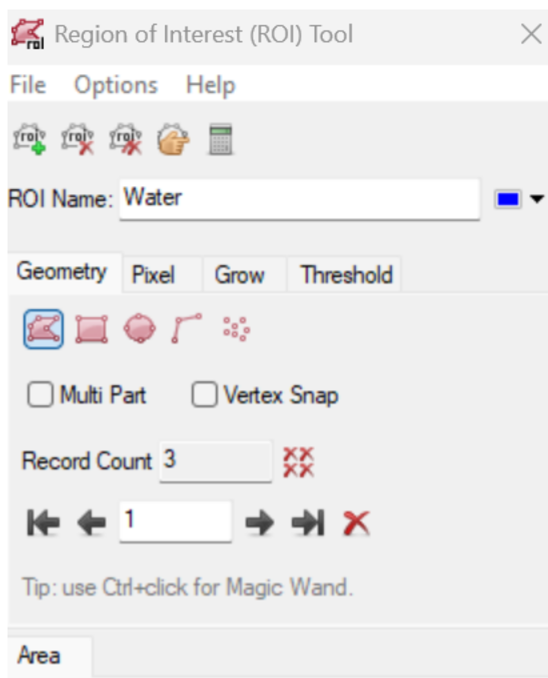


How to make your own ROI:

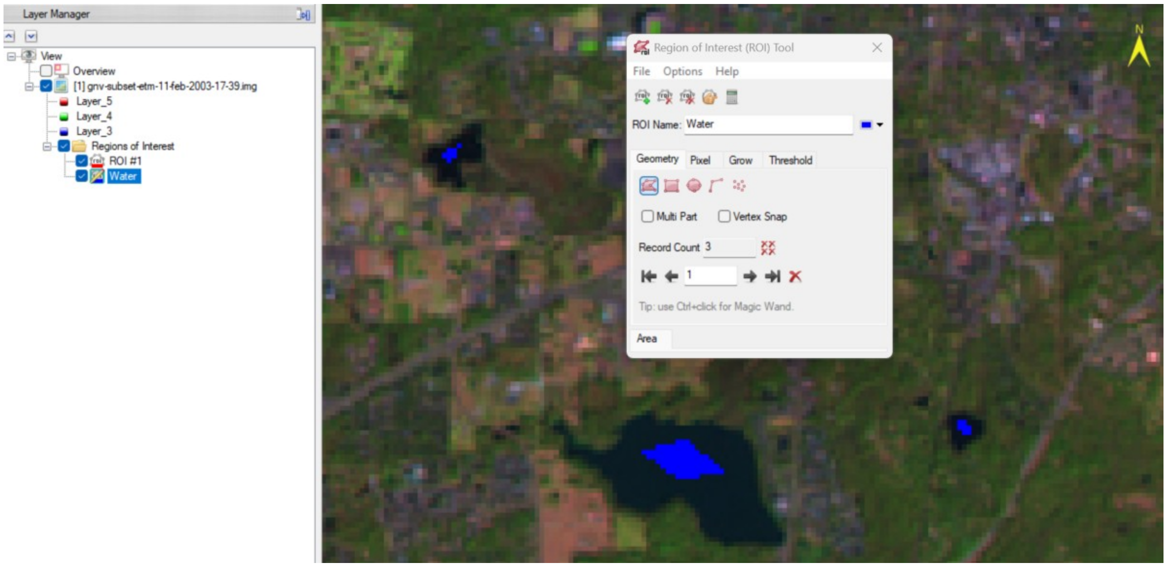
What if you don't have pre-existing training samples to create ROIs? in most of your work you will not have thousands of training samples. One solution is to manually create your own ROIs directly from the image (in addition to going to fieldwork and/or collecting samples from google earth).

1. Open ENVI and load the Gainesville ETM image (5,4,3).
2. Open the ROI tool from the top menu:

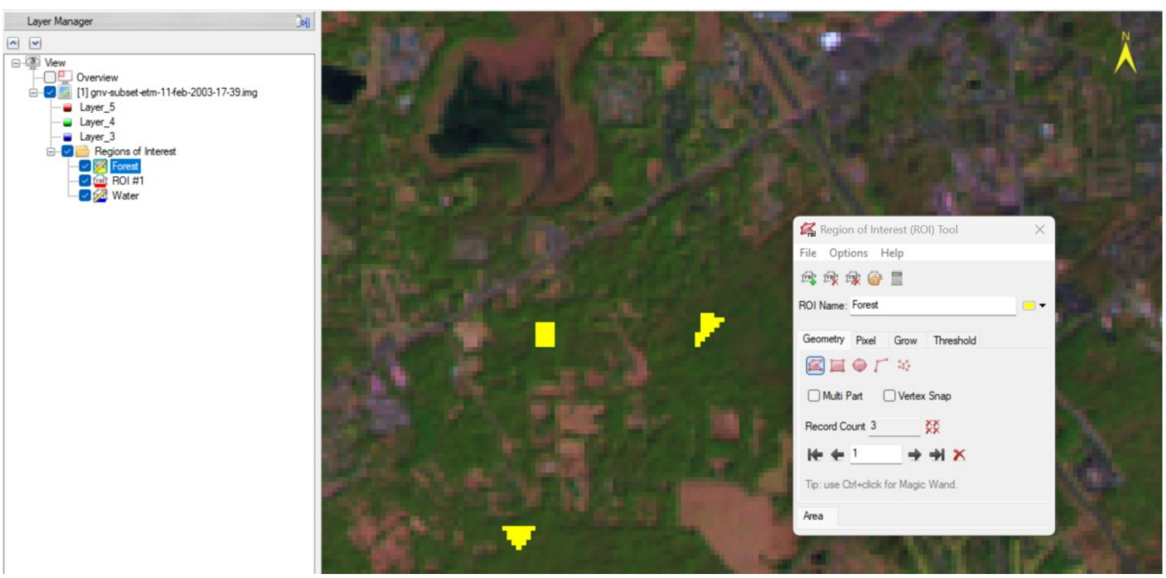
The Region of Interest dialog appears. Type in **Water** in the ROI Name field, leaving Geometry and Polygon as the defaults. Go to an area of image you think is a water body and draw a polygon there. You will draw the polygon and when the last line closes with the first point, a small circle appears. Right-click and accept **Complete and Accept** Polygon. Then go to other water bodies and draw a few polygon ROIs over other areas of water. You should try to include areas that encompass the variability of the class. Remember the issue of spatial autocorrelation as discussed in the lecture. In the ROI tool, also choose a color for water class, for example, blue.



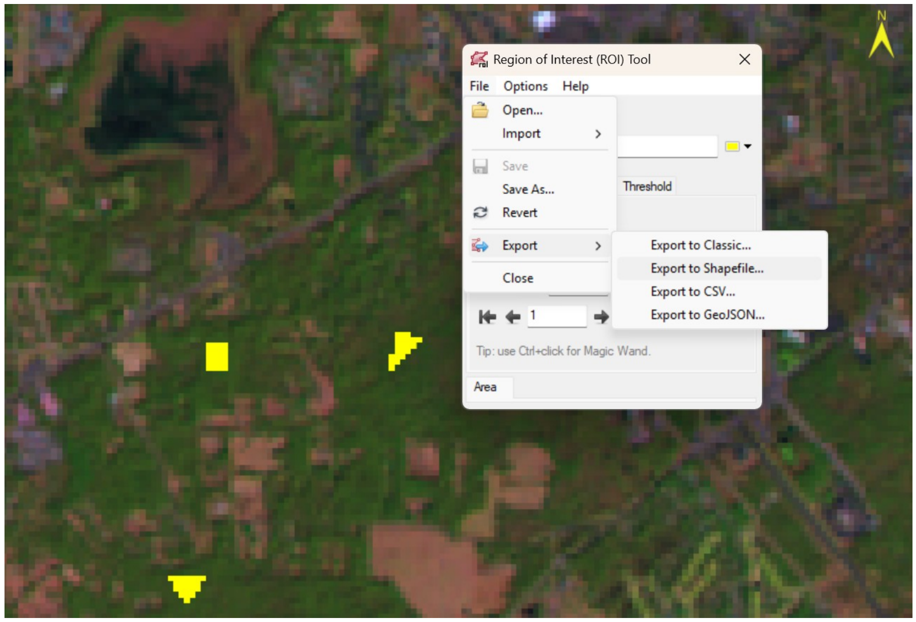
When you have enough Water ROIs (how many is enough? Remember the rules of thumb described in the lecture).



Similarly, add a new ROI and name it Forest, Urban, or whatever you want to make it next. Go through and create multiple ROIs for all of your land-cover classes. If you wish to stop for a while, right-click on the Regions of Interest in the Layer Manager and Save As… with an appropriate file name. You can reload it when you come back.



Finally, save the final ROI with all classes using the File menu> Export> Export to Shapefile so that you can use it in the future.



Note that you might also find yourself somewhere in between—not with thousands of training samples, but not entirely without them either. If you have only a limited number of training points, a good approach is to create larger polygons around each point (or around a selected subset of points) to serve as ROIs.

Well done on completing the exercise. Congratulations!