

GEO/EVS 425/525 Unit 11

Unsupervised Classification

In this laboratory, we continue our study of classification of satellite imagery. Recall that classification in general refers to the process of sorting pixels into a finite number of individual categories, based on their data file values. A pixel is considered to belong to a particular class if it corresponds to a set of criteria defining the class. The criteria are most commonly defined in one of two ways: via statistical clustering or via an estimate of likelihood based on prior identification of applicable classes. Last week, we carried out a supervised classification, based on prior identification of applicable classes. This week, we will carry out an unsupervised classification based on statistical clustering.

Unsupervised classification enables users to specify some basic parameters that the computer uses to uncover statistical patterns inherent in the data. The patterns do not necessarily correspond to meaningful characteristics of the area in question, such as soil type, vegetation type, or land use. But they are statistically closest to the distribution of the digital numbers in the raw image. The classes derived from the unsupervised classification are clusters of pixels with similar spectral characteristics. Their definition is dependent on the raw data. They have no inherent meaning. Therefore, it is the responsibility of the analyst to attach meaning to the classes. Unsupervised classification is usually used when relatively little is known about the data before classification. The classes resulting from the classification process have meaning only if the classes can be appropriately interpreted.

When an unsupervised classification is carried out, each class is associated with a range of variation of the digital numbers in the layers making up the raw image. Each range of variation associated with a particular class is termed a *signature*. The typical signature is based on a mean and a covariance matrix. These statistical parameters define a *parametric signature*. There are also nonparametric signatures, which are based on discrete objects such as polygons or rectangles, but we will not worry about nonparametric signatures in this exercise.

The classification process makes use of a *decision rule*, which assigns the data file values of the pixels in the signatures to specific classes. As with the signatures, decision rules can be parametric or nonparametric. A parametric decision rule necessarily assigns every pixel to a class, since parametric decision space is continuous; a nonparametric decision rule determines whether or not a pixel, based on its signature, lies within or outside a nonparametric signature boundary.

There is no single way to carry out an unsupervised classification. Two basic methods are in common use, and you will try both, using the TM image you rectified in Unit 9. The simpler uses three layers, identified with Red, Green, and Blue (RGB), respectively; it is called RGB clustering. The more complex can use as many layers as are found in the image; it is called the Iterative Self-Organizing Data Analysis technique (ISODATA).

RGB Clustering

RGB clustering is a relatively simple algorithm for classification and data compression for three bands of data. It is a fast and simple process that quickly compresses a 3-band image into a single-band pseudocolor image, without necessarily classifying any particular features into meaningful classes. The algorithm plots all pixels in 3-dimensional space and then partitions the space into clusters on a grid. There are two variants of the image-generation process. In the basic form, each of these clusters becomes a class in the output thematic raster layer. In the advanced form, assignments to class are based on a minimum threshold for belonging to a cluster. This allows for more color variation in the output file. Pixels that do not fall into any of the remaining clusters are assigned to the cluster with the smallest city-block distance from the pixel. In this context, *city-block distance* is calculated as the sum of the distance in the red, green, and blue directions in 3-dimensional space.

ISODATA Clustering

By contrast, with ISODATA, you can (and must) specify the number of layers to be used. You must also specify the number of clusters to be considered, the convergence threshold, and the maximum number of iterations to be performed. On the first iteration of the ISODATA algorithm, the means of the n clusters (where n is the number of clusters you specify) are determined arbitrarily. After each subsequent iteration, a new mean for each cluster is calculated, based on the actual spectral locations of the pixels in the cluster rather than the prior arbitrary (or more arbitrary) calculation. Then, these new means are used in defining clusters for the next iteration. The process continues until either the change from one iteration to the next is less than the convergence threshold you specified to begin the process or until the process has passed through the maximum number of iterations you specified. A convergence threshold of 0.95 indicates that processing will cease as soon as 95% or more of the pixels stay the same from one iteration to the next (or 5% or fewer pixels change). When you specify reasonable limits for convergence threshold and maximum number of iterations, you will insure that processing will not proceed forever. If the convergence-threshold limit does not kick in, the maximum-number-of-iterations limit will stop processing after a reasonable time.

Advantages and disadvantages of the two methods are as follows:

Advantages	Disadvantages
RGB Clustering Method	
The fastest method. It is defined to provide a fast, simple classification for applications that do not need specific classes	Exactly 3 bands must be input, which is not the most appropriate way to begin for all applications.
Not biased to the top or bottom of the data file. The order in which the pixels are examined does not influence the outcome	Does not always create thematic classes that can be analyzed successfully for informational purposes.
[Advanced version only] A highly interactive function, allowing iterative adjustment of parameters until the number of clusters and the thresholds are satisfactory for analysis	
ISODATA	
Clustering is not geographically biased to the top or bottom pixels in the file, since it is iterative	The clustering process is time-consuming, because it can repeat many times.
Highly successful at finding spectral clusters inherent in the data. It does not matter where the initial cluster means are located, as long as enough iterations are allowed	Does not account for pixel spatial homogeneity
A preliminary thematic raster is created, whose results are similar to a minimum distance classifier in supervised classification. This layer can be used for analyzing and manipulating signatures before the actual classification takes place.	

Clearly, both methods have both advantages and disadvantages, and neither is *a priori* preferable to the other. In this exercise, you will use the Remote-Sensed quadrangle you rectified in Unit 8 and classify it using both algorithms.

RGB Clustering

Open the Basic RGB Clustering dialog by clicking on Interpreter-GIS Analysis-RGB Clustering. Insert the name of your rectified TM quadrangle as the input and give the output a suitable name. Note the layers that will be used for the input. Note that the layers used, and their assignment to Red, Green, and Blue dimensions, are the same as the default layers for the viewer. Do you want this? 2-3-4 represents the most common assignment set for viewing; it is the same as the spectral balance for infrared film. Other useful assignments are 3-4-5, 2-4-7, 2-3-5, etc. If water is very important (and if the day was clear), you might also try 1-2-3. You might actually want to try several assignments (since this algorithm is fast) to see if it makes a difference. Leave the rest of the parameters at their default values. Click on OK to run the algorithm.

Look at the file you have just created. Does it appear to show classes that represent specific classes of land cover? Of anything else worth representing? Invoke ImageInfo to see the details of the image. There would appear to be 256 classes. Are there really? Look at the image histogram. Now click on Raster-Attributes to see what is really in the image. How many non-zero classes do you find? Take the most important ones, and give them contrasting colors (to do this, you might change all of the colors to black, and then change the colors with the largest histogram representation to different – and contrasting – colors. Knowing what you do about the area you have chosen, can you pick out any pattern? That is, is there a relationship between the classes shown on your map and the patterns that exist in the real world? Significantly, how many significant clusters are there in this image?

Now open the Advanced RGB Clustering dialog by clicking on Interpreter-GIS Analysis-Advanced RGB Clustering. Again, insert the name of your rectified TM quadrangle as the input file. Use the same bands assignment as you did in the basic classifier. Note the “state” window. It starts off saying “No data loaded into partitions.” Load the data by clicking on the “Load Image Data” button. The “state” window now says “Data partitioned; number of classes not calculated.” You have two choices, using the “Partition Data” tab: Autocalculate classes and Optimize the threshold for a number of classes you specify. You might try both. If you have the program calculate the number of classes, you will get a very large number of classes. Verify this for yourself, and run the algorithm for the large number of classes. Click on the “Image I/O” tab and insert a suitable name for an output file. Click on “Produce output.” The classified image is produced. Again, look at ImageInfo. All of the classes you had the computer calculate are present. Look at the image histogram. You will notice two things: all of the classes are represented by at least some pixels, and the pixels are strongly grouped into a series of clusters. Count the number of clusters.

Again open the Advanced RGB Clustering dialog and insert the name of your rectified TM quadrangle as the input file. Load the image data. You have now determined two possible optimal class numbers that should well be better than the very large number of clusters you derived from the totally automatic execution of the algorithm: the number of clusters you received from the basic RGB clustering model and the number of cluster groups you saw in the totally automatic calculation. Try running the model again using both of these numbers. What do you find?

Based on your knowledge of the area you have chosen, you should have some understanding of the distribution of land-cover types in your area: residential areas, parks, woodlands, commercial areas, pavement, etc. Which of the images best represents what you feel is there? Choose that one. Load it into your viewer. Click on Raster-Attributes. Then click on Edit-Add Class Names. A new column opens in the raster attribute table. Add those class names you feel some certainty about. Don't worry at this point if there is duplication. **This image, with its legend, should be included in your portfolio for this unit.**

ISODATA Clustering

You can begin an ISODATA clustering in either of two ways – but they are different. The simpler is to click on DataPrep-Unsupervised Classification; the alternative is to click on Classifier-Unsupervised

Classification. What is the difference in the dialogs that appear? For this round, it doesn't make a difference. It may for future exercises, however. Give the output file a suitable name. Choose a suitable number of classes. For a typical land-cover map, 10 should be enough. Consider the choices you made above in making your final choice. Then choose suitable numbers for Maximum Iterations and Convergence Thresholds. 25 should be enough iterations; 0.95 is a reasonable convergence threshold for most purposes. Click OK to start processing.

As with the image produced by your RGB clustering, you should have some understanding of the distribution of land-cover types in your area: residential areas, parks, woodlands, commercial areas, pavement, etc. Look at the image produced by your ISODATA clustering in a viewer. Click on Raster-Attributes. Then click on Edit-Add Class Names. A new column opens in the raster attribute table. Add those class names you feel some certainty about. Don't worry at this point if there is duplication. **This image, with its legend, should be included in your portfolio for this unit.**

Questions to Consider

1. For the RGB classification, how do you choose the most appropriate bands on which to base your classification?
2. Based on your unsupervised classifications, how would you decide which of these two approaches to unsupervised classification provides the best results? (What do you mean by "best results"?)
3. Again, based on your experience (and trial and error) with these two classification models, how do you choose the most appropriate number of classes in each case?
4. In answering question 2, do you know enough to make a choice? If not, what more would you like to know?

Portfolio

1. A classified TM image, based on your best RGB classification of the quadrangle you rectified in Unit 9, with legend.
2. A classified TM image, based on your best ISODATA classification of the quadrangle you rectified in Unit 9, with legend.