GEOG 213 –Intermediate Remote Sensing
Spring 2002 - Final Project

The use of exploratory clustering techniques in combination with the supervised classification of the West End of Santa Catalina Island

Marco Ruocco

Introduction

Image classification is a fundamental problem that in remote sensing can be approached in a variety of different ways. In particular, supervised and unsupervised techniques constitute the main strategies used by analysts to extract useful information from raw data. However, the same techniques used for unsupervised image classification can be used in an exploratory mode not to simply conduct new classifications but rather to expand our knowledge about the information classes captured by supervised techniques. One characteristic of exploratory clustering is in fact the ability to uncover the inherent data structures (Richards 1993) of information classes. The main purpose of this project is in to apply this concept of knowledge discovery to the information classes determined by supervised techniques for the West End of Santa Catalina Island.

This study continues the investigation of the land cover classification of the West End of Santa Catalina Island that was conducted by the author in the Geography 115B class in Winter 2002 using supervised classification techniques, documented in the corresponding project report. The attempt is to gain a deeper insight into the internal structure of supervised information classes determining aspects like variability and heterogeneity as they are exposed by different clustering parameters. Considerations about the importance and the limitations of the use of methods of qualitative visual inspection are provided in the conclusive remarks.

The data used in the study were three SPOT bands (Green, Red, and Near-Infrared) taken on March 15, 1990 and characterized by 20 m spatial resolution and 8-bit radiometric resolution. I chose the specific study area in the West End of Catalina

because it provided a representative sample of the variability of the entire island, at a smaller and more manageable scale.

Figure 1 is the Location Map for Santa Catalina Island, which includes terrain visualizations of the island and a false-color composite of the three SPOT images, with an inset map for the West-End.

Methods

The three SPOT bands were originally stored in Arc/GRID format. They were converted first in 24-bit true color BMP format using Arc Toolbox, and then into 8-bit grayscale IPW format using XV. All bands were preprocessed to remove the atmospheric effect using a simple offset method, consisting in subtracting from the DN values of the image the minimum DN value for that band, so that the darkest DN value was set to zero and all DNs were shifted accordingly.

A 500x350 pixels subset of the three SPOT images was selected to include the West End of Catalina and the immediate surroundings (the ocean and the isthmus of Two Harbors). The images were then cropped using XV along the specified boundary, generating the final preprocessed images used in the subsequent analyses. The advantage of this approach is that the automated classification is carried out on a much smaller area than the entire island, thus addressing with more efficiency the variability of the data.

Clustering is an unsupervised method that allows the classification of remote sensing images without *a priori* knowledge of the characteristics of the land cover classes. Clusters are a group of pixels sharing similar spectral characteristics, and they can be automatically determined from images using several different strategies and algorithms. The characteristics components of these procedures include the strategy employed (e.g. iterative or single pass), the method of determining the initial characteristics of clusters (i.e. seeding), the criterion that determines the degree of similarity between pixels (e.g. Euclidean or Manhattan distance), and the procedure of obtaining final clusters through processing (e.g. merging and splitting), among others.

Those aspects are reviewed in detail in Mather (1987), Jensen (1996), and Richards (1993), but in this study it is useful to focus only on a subset of them. In order to define such subset of interest the clustering algorithm "*cluster*" " implemented in IPW can be used as a reference. The algorithm performs an unsupervised cluster-based analysis of the input images and outputs a classified image along with multivariate statistics. The options available for this algorithm are (IPW):

- C: number of output classes, indicating the number of clusters generated and contained in the final output image.
- R: cluster threshold radius, the distance that determines the maximum extent in spectral space of each cluster. Also considered a measure of cluster compactness (Mather 1987).
- I: number of intermediate clusters retained during clustering. In practice this parameter determines when a pixel is used to form a new cluster kernel, on

the basis of the number of most populous clusters considered (only n = I clusters are considered in the process of classification).

- E: pixel exclusion value, that specifies the required pixel value in all bands for which the classification of that pixel will not be performed

The algorithm first chooses a randomly distributed set of cluster seeds, then merges the clusters according to a measure of similarity based on Euclidean distance, without splitting the clusters in case they have high variance. The particular clustering process in IPW applies the algorithm to all the pixels in the image and generates cluster statistics. Then the program uses the IPW program "bayes" to perform a maximum likelihood classification while considering the clusters just generated as separate classes. Finally the clusters are colored for image display (Geog213).

Among the parameters of the algorithm those of special interest here are the radius and the number of clusters. The radius parameter determines how large is the maximum extent of each cluster, and therefore controls the total variance, how subtle is the difference between clusters and how fine is the resolution of the partitioning of the spectral space.

The parameter controlling the number of clusters determines how many partitions are made in total in the spectral space. An attempt was made in this study to consider the outcomes of different radii and different cluster numbers. Since the two parameters control, albeit in different ways, the very same cluster variance measure, together with the procedure used to assign pixels to clusters, I found that filling a simple matrix of combinations (i.e. high cluster and low radius, medium cluster and low radius, etc.) was considered an unsuitable strategy. Instead some supposedly meaningful cases have been identified, investigating the outcomes of a stepped increase in the number of clusters while considering variations in a low radius range.

In order to analyze different clustering combinations the spatial extent of the supervised classes was compared visually to the spatial extent of the generated clusters. The clusters have in turn been classified and color-coded through a process of density slicing. The process assigned different color hues (red, green and blue) to different cover types (respectively bare ground, grassland and chaparral), and different color values (from light to dark in roughly uniform value steps) to the different instances of the three main land cover types generated by the cluster classification. In so doing it was possible to see at a glance the similarities of the cluster map with the supervised classification map (color hue), and the additional discrimination introduced by clustering (color value). One possible drawback of this visual classification was that the assignment of a cluster to a class was done visually and qualitatively, thus causing difficulties in boundary cases despite the simplification due to the broadness of the classes. For each cluster of each map the spectral signatures (main vectors) were displayed with the utility "liststats" and subsequently plotted.

The first step of the analysis was to generate the particular cluster map that was judged most similar to the supervised classification map. Other combinations could then be considered to explore further the characteristics of the classes. This approach is justified by the need for sharing a common classification base between supervised and

unsupervised techniques, from which the additional cluster-based analysis could be conducted. The assumption here was that the supervised classification provided a broad generalization of the land cover that could be further differentiated in different levels of clustering (each one specified by the number of clusters), in a hierarchical fashion.

Results

Among the several cluster maps that were generated, the following were selected for further inspection:

1) Clusters = 8, Radius = 4 (Figure 2B)
2) Clusters = 4, Radius = 8 (Figure 3A)
3) Clusters = 12, Radius = 8 (Figure 3C)
4) Clusters = 16, Radius = 8 (Figure 4A)
5) Clusters = 20, Radius = 8 (Figure 5A)
6) Clusters = 9, Radius = 4 (Figure 5C)

The first combination was judged to be the visually most similar to the supervised classification. Several aspects supported such evaluation:

1) The presence of a striking difference in land cover between south facing and north facing slopes, presenting respectively chaparral and grassland land cover types. This basic pattern emerges on any cluster map at all ranges of number of clusters, from 4 (Figure 3A) to 20 (Figure 5A), and on different radii, from 4 (Figure 5C) to 8 (Figures 2B, 3A, 3C, 4A, and 5A).
2) The presence of an additional pattern constituted by a coastal class that reflects the class of bare ground obtained from supervised classification. Very interestingly, however, the very similar clustering parameters (Clusters = 9 Radius = 4) of Figure 5C do not result in an analogous coastal class. Probably some sort of redistribution of variance has caused the more homogeneous look of grassland in this latter example.
3) The generation of a total of 3 chaparral clusters in the place of the supervised chaparral class, which accounted for both the classified and unclassified pixels of the supervised map. In a way the clusters do not add information to the supervised classification besides confirming the preexisting chaparral pattern and covering previously unclassified areas.

Figure 3A shows a minimal case with only two chaparral clusters and one grassland cluster, constituting a first example of chaparral/grassland differentiation in clustering. Figure 4C shows a cluster configuration with clusters = 12 and radius = 8 including 4 chaparral, 2 grassland and 4 bare ground clusters. Figure 4A is a configuration with clusters = 16 and radius = 8, with 7 chaparral, 2 grassland and 4 bare ground clusters. It shows a clearer breakdown of the rather uniform grassland cluster (considering lower clusters numbers) and its differentiation in fragments located at the

interface with the chaparral class and probably representing more healthy vegetation spots. Finally Figure 5A shows a configuration with clusters = 20 and radius = 8 containing 6 chaparral, 3 grassland and 6 bare ground clusters, which does not present a substantially different differentiation but rather a higher fragmentation and a high number of minor components.

Regarding the spectra, Figure 2D shows that within the preferred cluster map the land clusters, although clearly distinguished from the ocean classes, are themselves very similar to each other, showing in particular a typical healthy vegetation response that match only the spectral signature of the chaparral class of supervised classification in Figure 2B. Thus the clusters do not show the same differentiation as the supervised classes: in other words the spectral differentiation does not follow the visual differentiation.

This general pattern changes slightly for the other cluster combinations of Figure 3B, 3D, 4B, 5B and 5D where there is a clearer differentiation between the spectral response of the grassland clusters and that of chaparral classes, the former being characterized in general by a higher response than chaparral in all three bands, and in particular in the red band, probably due to the presence of a subcomponent of bare ground pixels. When considering the bare ground spectra they appear very similar to those of chaparral, having relatively low red DN values and a peak in infrared. Importantly they are very dissimilar from the spectral response of the supervised class of bare ground in Figure 2a.

Discussion

The preliminary investigation on the use of the cluster parameters of cluster radius and number of clusters involved an assessment of the effects of low radii in the classification of the image. A low radius may have in general similar results as a high cluster number, but particular situations may arise. For example, a low radius causes an image of the West End to be classified with many ocean classes, while a high number of clusters may result in only one or a few ocean classes. In other words, the two parameters control not only the variance of the output clusters (one directly in terms of maximum threshold value, the other indirectly, by the factor by which the total image variance is subdivided), but also the way that clusters are identified as such: it seems that in the case of the many ocean classes the available number of clusters is spent to capture the variability of the ocean, which in reality is a very minor part of the total variance of the image. This aspect seems to indicate the relevance and the difficulty of a process of reverse-engineering of the cluster algorithm that goes beyond the scope of this project. However it can be observed that the radius might determine not only the maximum extent of a cluster, but also causes the partitioning of the spectral space in ways that may not be representative of the actual distribution of image variance.

The results of the cluster classification offer an interesting insight into the characteristics of the original supervised vegetation classes and in particular in their level of internal heterogeneity.

First, the grassland class is represented by no more than two clusters in any of the considered combinations of cluster number and radius. This can be interpreted as a sign of high internal homogeneity, visually confirmed by the uniform cluster that covers most of the south part of the West End. Second, the breakdown of the uniform chaparral class in secondary clusters is achieved already at combinations of low cluster number and relatively high radius, indicating high internal heterogeneity that results in a fragmentation of its spectral characteristics. Third, the coastal bare ground cluster constitutes a secondary spectral feature that appears only with low radius or high cluster number. Nonetheless it is an existent effect that at the stage of supervised classification was detected thanks to the particularity of the training site.

The conclusions of the visual comparison of the classified maps in Figure 3 is not completely backed up by the comparison of the spectral signatures (mean vectors) of the clusters with those of the classes in the supervised classification. This would in turn suggest a visual classification problem or a particular way of how the clusters were generated, possibly incorporating pixels formerly distributed across the supervised classes. However the clear differentiation between chaparral and grassland classes confirmed the labels assigned to clusters and the overall interpretation of the landscape, except for the bare ground class. Concerning this latter point it might be observed that the bare ground class was a spectral peculiarity of the supervised classification that the clustering procedure processed in a different way, namely producing clusters with characteristics similar to the dominant chaparral clusters, although distinguished from them (see the spectral signatures differentiated in DN offset but not in structure).

For a physical interpretation of the phenomena shown by the cluster analysis the March 15th date of the image must be taken into account. The landscape was considerably influenced by peak seasonal precipitation, which caused grassland to reach its healthiest condition (confirmed by the highest spectral values). This in turn would influence the extent of the vegetation cover and reduce the differentiation between bare ground and grassland land covers. A second aspect to consider is the influence of the lighting geometry: the internal homogeneity of a great part of grassland might be caused by the homogenizing effect of sunlight on DN values, compared to the variability introduced by shadows in the north side (mainly chaparral) of the West End.

These factors combined suggest that the evaluation of the homogeneity of clusters and supervised classes might not have a straightforward physical correspondence, but rather they suggest the exaggeration of patterns in the landscape. However, from a spectral point of view, the grassland class, when isolated from the other extreme of its continuum, bare ground, is less variable than chaparral, and the large uniform (visually not shaded) green areas in all cluster maps seem to support the idea.

Conclusions

Exploratory cluster analysis can expand our knowledge about information classes developed through supervised classification techniques. The two strategies can

therefore be seen not in antithesis, but rather complementing each other. What supervised classification does not provide is an assessment of how variable are the identified classes internally and how easily the information contained in them can be broken down in subclasses in a hierarchical fashion. The parametric nature of cluster analysis allows us to derive such a hierarchy and test the inherent structure of a class by just considering the way it is broken down by an increasing number of clusters and/or by a decreasing threshold radius.

Grassland, when separated from bare ground, is a very homogeneous class, broken down only with high cluster number or very low radius. This apparently contradicts the idea presented in the project of Geog115B according to which the Bare Ground - Grassland formed a continuum on the southern slopes of West End. However it seems that such continuum is in reality non linear, since there is the majority of grassland that is homogeneous, but there is also a smaller transitional zone to Bare Ground and another interfacing with Chaparral. Cluster analysis, in other words, has provided us with information on the shape of the continuum that supervised analysis alone could not provide.

Conversely, Chaparral is a very heterogeneous class and in reality the term should be replaced by a more generic term since the component species, not investigated here directly, should be included in the general picture. Chaparral already comes with two clusters at a very low cluster number, and the number increases as the total clusters increase. No homogeneous sub-components seem to emerge, but, rather, a high degree of fragmentation characterizes this spectrally broad class.

In summary the idea that I am proposing for the analysis of the West End of Catalina is a hierarchical subdivision of classes, whereby the top two entities are Grassland and Chaparral, that in turn are broken down unevenly: Grassland in a homogeneous pattern of subclasses with only a minority of transitional types, while Chaparral has a more elaborated breakdown in a range of subclasses. The Cluster analysis has allowed us to explore the hierarchy, and the conceptual movement up and down the hierarchy is possible by utilizing the parameters offered by IPW of radius and number of clusters, which have been discussed here in some detail.

The cluster information expands the supervised information but the two are not in perfect agreement. In turn the two classifications might present a partial view on the physical reality. The challenge now is to consider the ground truth and provide physical explanations for the spectral behavior of the three information classes.

References

Geog213 *"Lab 7 - Clustering and data visualization"*, Intermediate Remote Sensing (Graduate course), Spring 2002, Department of Geography, University of California at Santa Barbara. Instructor: Bill Bushing.

IPW (Image Processing Workbench) User Manual, http://www.geog.ucsb.edu/~geog115/

Jensen, J.R. (1996) "*Introductory digital image processing*", Prentice Hall, Upper Saddle River, New Jersey.

Mather, P.M. (1987) "*Computer processing of remotely sensed images: an introduction*", John Wiley & Sons.

Richards, J.A. (1993) "*Remote sensing digital image analysis: an introduction*", Springer-Verlag, Berlin.