Applying Supervised Clustering to Landsat MSS Images into GIS-Application

M. Torres, M. Moreno, R. Quintero, and G. Guzmán

Geoprocessing Laboratory-Centre for Computing Research-National Polytechnic Institute, Mexico City, Mexico
{mtorres, marcomoreno, quintero, sergei}@cic.ipn.mx
http://geo.cic.ipn.mx

Abstract. In this paper, we propose and implement an algorithm to perform a supervised clustering classification for geo-images. The *Maximum Likelihood Classification* method is used to generate raster digital thematic maps by means of a supervised clustering. The clustering method has been proved in Landsat MSS images of different regions of Mexico to detect several training data related to the geographic environment. The algorithm has been integrated into Spatial Analyzer Module to improve the decision making model and the spatial analysis into GIS-applications. On the other hand, the GIS-application contains a spatial database to store the vector and raster data, which are used in the spatial analysis process.

Keywords: Supervised Clustering Method, Geographical Information System, Maximum Likelihood Algorithm, Spatial Analyzer Module.

1 Introduction

The integration of remote sensing and geographic information systems (GIS) in environmental applications has become increasingly common in recent years. Remotely sensed images are an important data source for environmental GIS-applications, and conversely GIS capabilities are being used to improve image analysis procedures. In this case, when image processing and GIS facilities are combined in an integrated system vector data, they can be used to assist in image classification and raster image statistics, within vectors can be used as criteria for vector query and analysis [1].

Nowadays, the geographic information plays an important role in the decision making process, the spatial databases are very useful and powerful tools to handle, display and process spatial data. Frequently the need arises to analyze mixed spatial data. The data sets can consist of satellite spectral, topographic and other point form data, all registered geometrically, as might be found in a GIS [2].

In this paper, we propose a supervised clustering algorithm applied to Landsat MSS images. The classification method is used to detect *training data* according to the characteristics of particular Landsat MSS images. This method has been integrated into GIS-Application. Our approach can be considered as a part of spatial analysis, because the method is focused on processing raster spatial data to find out new properties, which can define the behavior of a certain geographic environment.

Moreover, these properties can be considered as spatial semantics to know the characteristics, which are not located with common approaches.

The rest of the paper is organized as follows. In section 2 we present the architecture of the GIS-application. The supervised clustering method and algorithm applied to Landsat MSS satellite images are described in section 3. Section 4 presents some results obtained by the GIS-application. Our conclusions are sketched out in section 5.

2 Architecture of GIS-application

The GIS-application has been developed using client-server architecture. This tool contains the following components:

Spatial Database (SDB). This module stores the spatial data (vector and raster) into a hierarchical structure. SDB contains a spatial dynamic index mechanism to organize physically the geographical objects according to the basic primitive of representation [3].

ArcMap GUI. This mechanism is a common wateway, which is used to process the requests of the users. The results obtained by the spatial analysis are rendered in this component.

Administration Module (AM). AM is used to control all the processes of the GIS-application. This module receives all the requests that the users have generated to perform any spatial analysis.

Spatial Analyzer Module (SAM). This module has been designed to make spatial analysis procedures. It includes the *supervised clustering method* to identify the characteristics of the raster data. SAM is composed by several methods to make spatial and visual analysis with the geo-information such as flooding and landslide detection areas.

The functional mechanism of the GIS-application is the following: ArcMap users make a request. This request is sent by DCOM technology to the AM to interact with the Enterprise GIS. AM processes the request and sends the parameters to SAM. In the Enterprise GIS, it is necessary to verify the definition to obtain the spatial data from SDB. The geographical objects are stored in the spatial database in which they will be analyzed by SAM. SAM is focused on detecting the characteristics of vector and raster data to detect flooding and landslide areas. Fig. 1 depicts the architecture of the GIS-application.

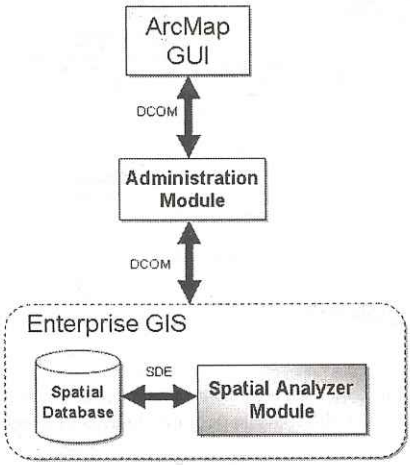


Fig. 1. Architecture of GIS-application

3 Supervised Clustering Method

3.1 Supervised Clustering

The supervised clustering is the procedure used for quantitative analysis of remote sensing image data. It rests upon using suitable algorithms to label the pixels in an image as representing particular ground cover types, or classes [4]. In this work, we have implemented the *Maximum Likelihood Classification* method. For supervised clustering, training regions are defined and pixels are assigned to one of these a priori defined classes.

There are five basic steps that we followed for supervised clustering: (I) Determine the number and type of classes to use for the analysis, (II) Choose training regions for the classes to identify the spectral characteristics typical for each specific class, (III) Use these training regions to determine the parameters of the supervised clustering, (IV) Then classify all image pixels, assigning them to one of the defined classes by the training regions, and (V) Summarize the results of the supervised clustering [5].

An important assumption in supervised clustering usually adopted in remote sensing is that each spectral class can be described by a probability distribution in multispectral space. A distribution describes the chance of finding a pixel belonging to that class at any given location in multispectral space. A two dimensional multispectral space with the spectral classes is modeled in Fig. 2.

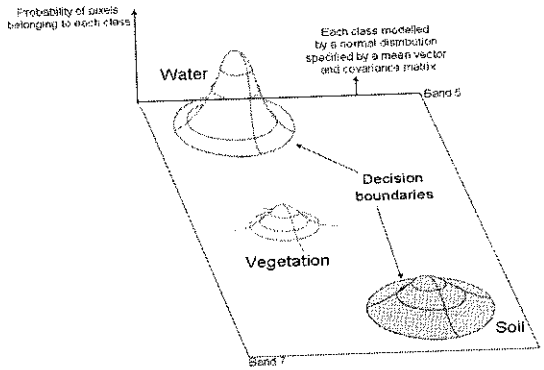


Fig. 2. Two dimensional multispectral space with the spectral classes represented by Gaussian probability distributions. Here the space is defined in terms of Landsat MSS bands 5 and 7

The decision boundaries shown in Fig. 2 represent points in multispectral space, where a pixel has equal chance of belonging to two classes. A multidimensional normal distribution is described as a function of a vector location in multispectral space by:

$$p(x) = \frac{1}{(2\pi)^{N/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(x-m)^{t}\Sigma^{-1}(x-m)\right\}, \quad (1)$$

where:
x is a vector location in the N dimensional pixel space.
m is the mean position of the spectral class.
$\sum$ is the covariance matrix of the distribution, which describes its spread directionally in the pixel space.

The multidimensional normal distribution is completely specified by its *mean vector* and its *covariance matrix*. Consequently, if the mean vectors and covariance matrices are known for each spectral class, then it is possible to compute the set of probabilities that describe the relative likelihoods of a pattern at a particular location belonging to each of these classes. It can be considered as belonging to the class, which indicates the highest probability. Therefore it m and $\sum$ are known for every spectral class in an image, every pixel can be examined and labeled corresponding to the most likely class on the basis of the probabilities computed for the particular location for a pixel. Before the classification, m and $\sum$ are estimated for each class from a representative set of pixels, commonly called a *training set*.

3.2 Supervised Clustering Algorithm

In this work, we implement a Maximum Likelihood Classification method and we have considered the following properties:

Let the spectral classes for an image that is represented by (2):

$$\omega_i, i = 1,...,M, \tag{2}$$

where M is the total number of classes. By attempting to determine the class or category to which a pixel at a location x belongs to its strictly the conditional probabilities:

$$p(\omega_i \mid x), i = 1,...,M \tag{3}$$

The position vector x is a column vector of brightness values for the pixel. It describes the pixel as a point in multispectral space with coordinates defined by the brightness, as shown in Fig. 2. The probability $p(\omega_i \mid x)$ provides the likelihood that the correct class is ω_i for a pixel at position x. Classification is made-up according to:

$$x \in \omega_i \text{ if } p(\omega_i \mid x) > p(\omega_j \mid x) \text{ for all } j \neq i \tag{4}$$

i.e., the pixel at x belongs to class ω_i if $p(\omega_i \mid x)$ is the largest. This intuitive *decision rule* is a special case of more general rule in which the decisions can be biased according to different degrees of significance being attached to different incorrect classification.

Suppose now that sufficient training data is available for each ground cover type. This can be used to estimate a probability distribution for a cover type that describes the chance of finding a pixel from class ω_i, and at the position x. Later the form of this distribution function will be performed more specific. For the moment, it will be retained in general terms and represented by the symbol $p(x \mid \omega_i)$. There will be as many $p(x \mid \omega_i)$ as there are ground cover classes. On the other hand, for a pixel at a position x in multispectral space, a set of probabilities can be computed that provide the relative likelihoods that the pixel belongs to each available class. The desired $p(\omega_i \mid x)$ in (4) and the available $p(x \mid \omega_i)$ estimated from training data are related by Bayes' theorem, which is shown in (5):

$$p(\omega_i \mid x) = p(x \mid \omega_i) p(\omega_i) / p(x), \tag{5}$$

where $p(\omega_i)$ is the probability that class ω_i occurs in the image, $p(x)$ in (5) is the probability of finding a pixel from any class at location x. We can note that:

$$p(x) = \sum_{i=1}^{M} p(x \mid \omega_i) p(\omega_i) \tag{6}$$

Although $p(x)$ itself is not important in the calculus. The $p(\omega_i)$ are called *a priori* probabilities, since they are the probabilities with which class membership of a pixel could be guessed before classification. By the comparison the $p(\omega_i \mid x)$ are posterior probabilities. Using (5) it can be seen that the classification decision rule of (4) is:

$$x \in \omega_i \text{ if } p(x \mid \omega_i) p(\omega_i) > p(x \mid \omega_j) p(\omega_j) \text{ for all } j \neq i, \tag{7}$$

where $p(x)$ has been removed as a common factor. The rule of (7) is more acceptable than (4) since the $p(x)$ are known by training data, and it is conceivable that the $p(\omega_i)$ are also known or can be estimated from the analyst's knowledge of the image.

In maximum likelihood classification, training regions are used to estimate the mean and covariance for each class. In this analysis, we assume that the classes have multidimensional normal distributions and that to each image pixel can assign a probability of being a member of each class. After computing the probabilities of a pixel being in each of the available classes, we assign the class for which it has the highest probability. The analysis produces a *Class Distribution Layer*, a layer contains the computed probability for each pixel for each class, and a layer provides a typicality index for each pixel class.

All of these layers are used to determine the effectiveness of our training regions and the subsequent classification analysis [5]. The supervised clustering algorithm consists of the following steps:

[Step 1]. Determine the number of ω_i for the classification (training fields).

[Step 2]. Select the pixel that covers the spectral quantitative value by means of a vector x, including the location.

[Step 3]. For each image pixel, determine a feature vector x by $p(\omega_i \mid x), i = 1,...,M;\ x \in \omega_i$.

[Step 4]. Assign labels to every ω_i.

[Step 5]. Update labels in each image pixel x_0, applying the current label vectors x_n and local spectral vector x to the decision rule $x \in \omega_i$ if $p(x \mid \omega_i)p(\omega_i) > p(x \mid \omega_j)p(\omega_j)$ for all $j \neq i$.

[Step 6]. Compute the maximum likelihood distribution and the covariance matrix for each class generated using equation (1).

[Step 7]. Finalize the algorithm if the labels of all pixels in the image are stable, repeat the second step otherwise.

4 Results

By using GIS-application, we can perform a supervised clustering in Landsat MSS images. The algorithm has been implemented into SAM [6]. The implementation has been made in C++ Builder to interact with the GIS-Application. Sufficient training samples for each spectral class should be available to allow reasonable estimations of the elements of the mean vector and the covariance matrix to be determined.

For an N dimensional multispectral space at least $N+1$ samples are required to avoid the covariance matrix being singular. As a result of the classification algorithm, the segment of Tamaulipas State Landsat MSS image is shown in Fig. 3. This is a 256 x 276 pixel array of image data in which four broad ground cover types are evident. These are water, fire burn, vegetation and "developed" land (urban). Also, Fig. 3 shows the locations of four *training fields* used to make the supervised clustering.

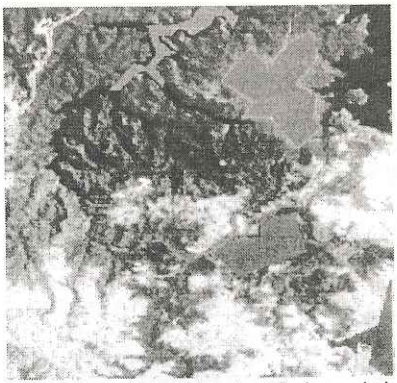


Fig. 3. Image segment of Tamaulipas State to be classified, consisting of a mixture of natural vegetation, waterways, urban development and vegetation damaged by fire. Four training regions are identified in solid color. They are water (magenta), vegetation (green), fire burn (red), and urban (blue). Pixels from these were used to generate the signatures in Table 1

We have considered that to obtain a good estimation of class statistics, it may be necessary to choose several trainings fields for the one cover type, located in different regions of the image. The four band signatures for the classes are obtained from the training fields, which are given in Table 1.

Table 1. Class signatures generated from the training areas in Fig. 3. Numbers are in the scale of 0 to 255 (8 bits)

Class	Mean vector	Covariance matrix			
Water	44.27	14.36	9.55	4.49	1.19
	28.82	9.55	10.51	3.71	1.11
	22.77	4.49	3.71	6.95	4.05
	13.89	1.19	1.11	4.05	7.65
Fire burn	42.85	9.38	10.51	12.30	11.00
	35.02	10.51	20.29	22.10	20.62
	35.96	12.30	22.10	32.68	27.78
	29.04	11.00	20.62	27.78	30.23
Vegetation	40.46	5.56	3.91	2.04	1.43
	30.92	3.91	7.46	1.96	0.56
	57.50	2.04	1.96	19.75	19.71
	57.68	1.43	0.56	19.71	29.27
Developed (urban)	63.14	43.58	46.42	7.99	-14.86
	60.44	46.42	60.57	17.38	-9.09
	81.84	7.99	17.38	67.41	67.57
	72.25	-14.86	-9.09	67.57	94.27

The mean vectors can be seen to agree generally with known spectral reflectance characteristics of the cover types. Also, the class variances (diagonal elements in the covariance matrices) are small for water as might be expected but on the large side for the developed (urban) class, indicative of its heterogeneous nature. Using these signatures in a maximum likelihood algorithm to classify the four bands of the image in Fig. 3, the thematic map depicted in Fig. 4 is obtained. The four classes by area are pro-

vided in Table 2. Note that there are no unclassified pixels, since a threshold was not used in the labeling process. The estimation areas are obtained multiplying the number of pixels per class by the effective area of a pixel. In the case of the Landsat MSS the pixel is 4424 m.

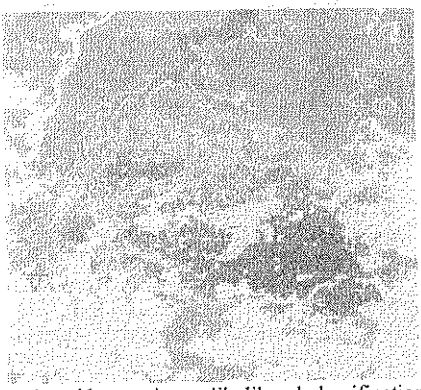
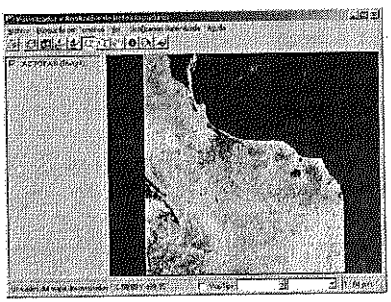


Fig. 4. Thematic map produced by maximum likelihood classification. Blue color represents water, red is fire damaged vegetation, green is natural vegetation and yellow is urban development

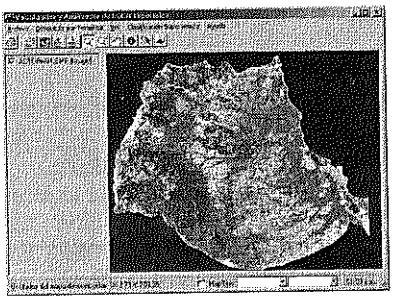
Table 2. Tabular summary of the thematic map of Fig. 4

Class	Number of pixels	Area (m)
Water	4830	21370000
Fire burn	14182	62740000
Vegetation	28853	127650000
Developed (urban)	22791	100830000
Developed (urban)	22791	100830000

Other result (Fig. 5 (a)) shows the Nautla Basin (Landsat) image, which is processed by GIS-application. The result of supervised clustering is shown in Fig. 5 (b).



(a) Landsat image of Nautla Basin



(b) Thematic map of Nautla Basin

Fig. 5. In (a) Landsat image is processed into GIS-Application; (b) Thematic map is generated by means of a supervised clustering algorithm

5 Conclusions

In the present work the supervised clustering method for Landsat images is proposed. The maximum likelihood classification algorithm recognizes patterns (training fields) to identify other pixels with similar characteristics. When applying this algorithm, it is essential to use *a priori knowledge* related to training fields, giving by human experience and several approaches of remote sensing.

This algorithm aims to preserve spectral properties and relationships between information and spectral classes. The method is used to make a quantitative analysis of sensing spatial data.

The algorithm basically consists of computing of the mean vector and covariance matrices for each spectral class and after that it is possible to compute the set of probabilities, which describe the relative likelihoods of a pattern at a particular location belonging to each of these classes.

The supervised clustering algorithm is essential for the decision making process. Moreover, it is used to improve the spatial analysis in different geographical environments. The approach allows us making more accurate in spatial analysis with data fusion (vector and raster).

This algorithm has been incorporated into GIS-application by means of the Spatial Analyzer Module, which can be incorporated to detect flooding and landslide areas. The algorithm has been implemented in *C++ Builder*, all the tests were performed in Landsat MSS geo-images that are composed of 5 spectral bands. The spatial resolution used in the process was 32 m. per pixel. The supervised clustering signatures are classified and integrated in a same render to improve the performance of visualization.

In addition, this approach can be used to know other useful spatial and attributive properties, which define the *spatial semantics* of geo-images

Acknowledgments

The author of this paper wishes to thank the Centre for Computing Research (CIC), General Coordination of Postgraduate Study and Research (CGEPI), National Polytechnic Institute (IPN) and the Mexican National Council for Science and Technology (CONACYT) for their support.

References

1. Hinton, J.C.: GIS and remote sensing integration for environmental applications. International Journal of Geographic Information System, 10, (1996) 877-890.
2. Ehlers, M.: Remote sensing and geographic information systems: towards integrated spatial information processing, IEEE Transactions on Geoscience and Remote Sensing, 28, (1998) 763-780.
3. Rigaux, P., Scholl, M., Voisard A.: Spatial Databases with Application to GIS. Morgan Kaufmann Publishers. United States of America (2002) 312-332.
4. Atkinson, P.M., Tate, N.J. (ed.): Advances in Remote Sensing and GIS Analysis. John Wiley & Sons. England (1999) 167-185.

5. Richards, J.A., Jia, X.: Remote Sensing Digital Image Analysis, An Introduction. 3rd edn. Springer-Verlag, Berlin Heidelberg New York (1999) 181–192.
6. Torres, M., Moreno, M., Menchaca, R., Levachkine, S.: Making Spatial Analysis with a Distributed Geographical Information System, Proceedings of IASTED International Conference on Databases and Applications. Acta Press, ISBN: 0-88986-341-5, ISSN: 1027-2666, Innsbruck, Austria (2003) 1245-1250.