URBAN AREAS IDENTIFICATION THROUGH CLUSTERING TRIALS AND NEURAL NETWORKS

14TH European Colloquium on Theoretical and Quantitative Geography
September 9-13, 2005, Tomar, Portugal

Júlia LOURENÇO, Luís RAMOS, Rui RAMOS, Henrique SANTOS and Delfim FERNANDES
UMINHO - UTAD

The main objective of this paper is to assess how an urban area can be identified accurately using satellite images. The case study is the urban centre of Vila Real and the satellite image used is SPOT5.

Results are still being worked out and some shortcomings of the process are known. A most important one is the lack of large data sets as only one satellite image is normally used. This problem is typical in urban studies due to the funding shortage. Important decisions are concerned with the number of classes and their homogeneity as well as the estimation of accuracy.

Following the obtained results, it will be possible to choose the most efficient classification in terms of performance, accuracy and confidence level upon working method. Likewise, the monitoring process of urban areas expansion can be used more extensively.

KEYWORDS

Remote Sensing, Neural Networks, Maximum Likelihood Classification, Supervised Classification.

INTRODUCTION

Satellite image processing techniques associated with thematic mapping allow the identification of the various land uses (urban, clear-cut areas, agriculture, forest land, water, wetland, etc.) being the urban areas the focus of the research.

The statistical models of classification require hypothesis at the level of the class distribution. This condition is not easily satisfied in several cases. Recent research has used techniques of artificial intelligence to recognise patterns and classification of satellite images.

This work presents techniques of maximum likelihood based on neural networks and compares their performance on image classification versus conventional methods. The final objective is to quantify the amount of urban area that can be traced in remote sense digital images. The relevance of this topic lies on the possibility to assess the validity in terms of accuracy of the classification.

A first methodological step is to choose the type of satellite image to be used. A second step is to select the appropriate classification techniques. The next subsections introduce the criteria for the choices undertaken at both steps. There follows a description of the workflow, a discussion of obtained results and concluding remarks.

SATELLITE IMAGES SELECTION

In this section a summary explanation of the main concepts associated with the production of maps derived from satellite images is presented. The way satellites obtain Earth surface images is briefly described in order to select the most adequate satellite. The sensors of Earth observation satellites capture the solar energy which is reflected by the objects in various areas of the electromagnetic spectrum. These areas are known as bands or channels and the number of bands of a satellite is named as spectral resolution. The value of one pixel in a band is a digital number (DN), representing the radiance conversion that reaches the sensor at an interval of full values from 0 to 255. All the bands are stored in raster format.

The images that are used more frequently on the production of land use maps are obtained by satellites SPOT/HRV and Landsat-TM, whose characteristics are described in table 1.

Table 1: Characteristics of the main satellites

Satellites/Sensors	Type of use	Spatial Resolution	Spectral Resolution
SPOT 5	Land-use	10 m * 10 m 20 m * 20 m	1 band 3 bands
Landsat/TM	Land-use	30 m * 30 m	6+1 bands

The reason why the satellites images can be used for the production of land use maps is related with the different form by which the several land objects reflect the solar energy. It is the quantity of reflected energy by the area that is converted in digital numbers for each pixel. These ND are subject to statistical analysis more or less sophisticated to convert the ND in land-use classes. The way this analysis is performed is described at a later stage.

Figure 1 shows the reflectance of the main surface elements which are water, vegetation and soil. The filtered bands of the main observation satellites were conceived to capture the radiance in the spectral zone in a way to promote an easy identification.

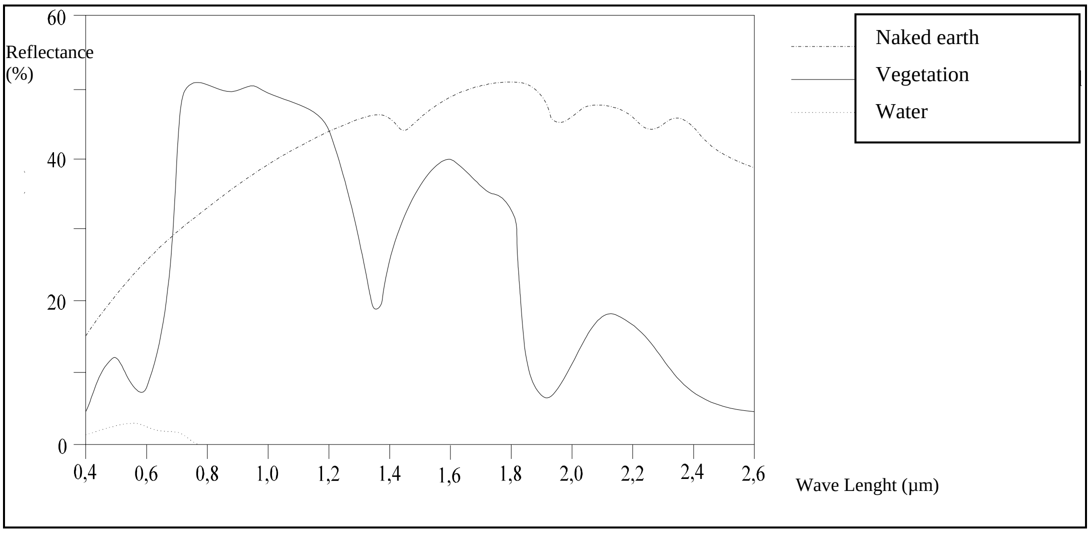


Figure 1: Reflectance of the main surface elements

CLASSIFICATION TECHNIQUES

Conventional Classification

Since most of image classifications are based solely on spectral signatures of land classes, the success of this operation will depend mostly on two aspects:

1 – The presence of different signatures to the various relevant land use classes;

2 – The possibility to distinguish with confidence these signatures from other non-relevant patterns.

A fundamental outcome of the classification process is the accuracy estimation of the final produced images. This involves the identification of a certain number of areas that must be confirmed through field survey which is normally a rather cumbersome task. Likewise, accurate statistical estimation for all study area as well as for the individual classes can be obtained.

Assisted Classification

This section deals with the concepts involved in assisted classification as undertaken at present research.

Three groups of such classification techniques can be found: statistical studies (parametric and non-parametric), neural networks and techniques of spectral mixes.

Typically [1], solving a classification problem consists on defining the existing relationships between three elements – classes/ objects/ observations, as described in Figure 2.

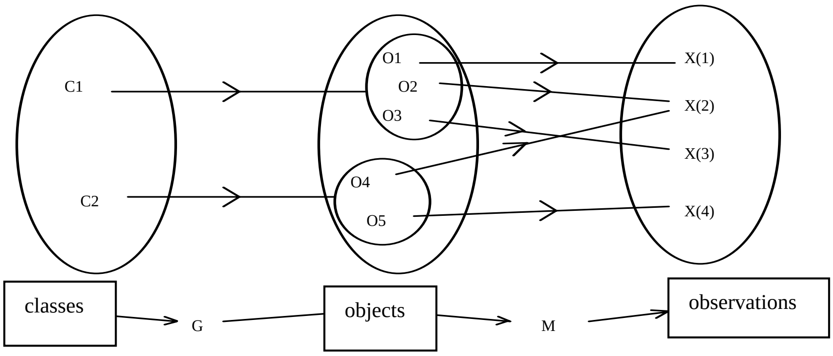


Figure 2: Relations admitted for the assisted classification

The first set, classes, is related to land-uses, as listed before and normally consists of a small number of elements.

The second set, objects, consists on the surface areas identified by each pixel pattern.

The third set, observations, are the characteristics observed from each object, such as the radiometric or topographic characteristics.

Concerning the relations between all these elements it is considered that each object belongs to one class only. That is, it is assumed the hypothesis of the “pure pixel” although it can also be assumed that pixels can belong in a certain degree to more than one class. This will be discussed when presenting the spectral mixes analysis.

The problem consists therefore in defining rules to establish the relationships between class, object and observations so that for a certain set of observations, an automatic classification can be obtained.

In order to find a solution for the problem, a training sample is required. This training sample is a sub-set used to define the rules to assign the new objects to the classes. Besides, a test sample is also required in order to evaluate the accuracy of the defined rules.

In short, there is a search for a decision rule that classifies with the utmost accuracy the objects, avoiding at the same time a large dependence of the used sample.

In the research being carried out at present, in a first instance, the classes that are the land-uses were defined at start and the sample was also predefined.

The classification method (MAXLIKE)

This classifier is the most frequently used and it is applied through the so-called MAXLIKE routine which is based on a probabilistic function associated with a signature obtained from a training sample. This methodology was adapted from Richards, 1986.

The pixels are grouped within the most probable class by comparison with probability of belonging to each of the signatures. This is the worst time-response classifier albeit the more accurate when good signatures are available. On the contrary, if that is not the case for the maximum likelihood method, the minimum distance classifier will be more appropriate. Additionally, as a general requirement, the number of pixels collected for the identification of a certain signature should not be lower than ten times the number of used bands.

This classifier allows the previous incorporation of the probability of one pixel belonging to a certain class, whenever this can be estimated. When this is not possible, equal probability is assumed for all classes.

The MAXLIKE routine allows also the possibility to exclude a given fraction of pixels with less probability of belonging to any other considered class. This excludes those pixels from being classified.

NEURAL NETWORKS

The proximity/ connectivity is a paradigm developed in the framework of artificial intelligence and that has had considerable success in the problem solving classification [4,6].

A neural network is a structure composed by interconnected processors. Some of these processors referred usually as nodes receive the *input* and others generate *output*. When a neural network is used for assisted classification, a set of examples are required to train the network, aiming to find an optimum weights set for the connections. The optimality is assessed with a criteria that compares for each element of the sample, the output of the network with the class that each element belongs in reality.

In engineering the most used neural networks belong to the *feedforward* type. In these networks the nodes are organised in layers, the first layer being the one for input and the last one for output. The number and dimension of the intermediary levels is variable. There are connections between one layer and the next layer and the connections are always established in the same direction. The architecture of the network influences the results and the choice of the best architecture for a certain problem does not have, in many cases, an obvious answer. The object attributes make up the input level and the class identification, the output level. These methods only admit the objects representation through $\Re^p$ vectors, and provide non explicit decision rules for the relevant concepts to describe the problem. But they have the advantage of preventing the need to use algorithms or rules in terms of the problem domain.

DATA MINING

The interest in the Knowledge Discovery from Databases (KDD) and Data Mining (DM) arenas arose due to the rapid emergence of electronic data management methods, leading to an explosive growth of business, government and scientific data bases [5].

The terms KDD and DM are often confused. KDD denotes the overall process of transforming raw data into high-level knowledge. It consists in a series of iterative steps such as [5]:

- understanding the application domain;

- acquiring or selecting a target data set;

- data cleaning, pre-processing and transformation;

- choosing the DM goals and learning algorithm;

- searching for patterns of interest;

- result interpretation and verification;

- using and maintaining the discovered knowledge.

Thus, DM is just one step of the KDD process, aiming at the extraction of useful patterns from observed data.

The DM approach incorporates in a first stage the choice of the objectives (for instance, forecasting, description and so on) taking into consideration the global process of Knowledge Discovery from data bases. This can be executed through the following tasks [7]:

- *classification* - labelling a data item into one of several predefined classes;
- *regression* - mapping a set of attributes into a real-value variable, that is a search for a mapping function between objects and classes;
- *clustering* - searching for natural groupings of objects based on similarity measures;
- *link analysis* - identifying useful associations between data.

A model is defined as a function (map) that attributes to each possible sample on the domain defined by the entry attributes, a value contained in the domain of the output attributes [2]. Each model contains a set of parameters that have to be adjusted or estimated after submitting a data set to an algorithm (learning stage).

On the validation stage one intends to have an estimation on the quality/ performance of the model [3]. The test sample allows model’s performance evaluation enabling its comparison with other learning techniques. A third and extra set is needed in case of adjustment of meta parameters (e.g., algorithm learning rate). The validation set allows the choice of the model with the better estimation error.

CASE STUDY AREA

The study area is located in Vila Real in Vila Real Municipality that belongs to the interior of the Northern Region of Portugal (see figure 3) in a peripheral area of Western Europe.

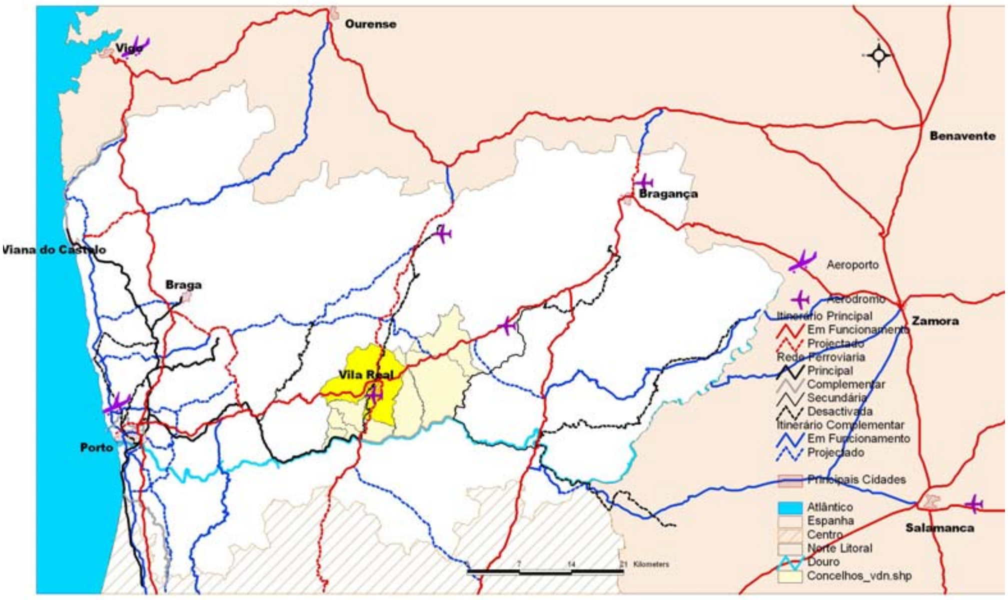


Figure 3: Location of study Area on the Northern Region of Portugal

Vila Real leads in area size and functional dimension a sub-regional district comprising seven Municipalities occupying 1203 Km^2 (see table 2)

Table 2 : The area under study

Municipalities	Area (km^2)	Area (%)	No. of Parishes
Alijó	292,8	23,7	19
Mesão Frio	26,1	2,6	7
Murça	191,3	15,8	9
Peso da Régua	89,9	7,9	12
Sabrosa	154,2	13,2	15
Santa Marta Penaguião	70,7	5,3	14
Vila Real	**377,7**	**31,6**	**30**
TOTAL	1202,7	100,0	106

Source:INE

A sample of the satellite image (see figure 4) was chosen for the analysis of the urban area of the town of Vila Real. This medium – sized town has had the highest population increase in the interior areas of Portugal. It is located in a hilly transition area between the coast and the interior of Portugal. This town had a big family and accommodation dynamics after the eighties due to the expansion and up-grading of its higher education equipment: from higher studies Institute to University.

Road infrastructure development has been lagging behind hindering the north-south connection axis between Chaves in the Northern border with Spain and Régua/Lamego, two towns on Douro river banks.

The Axis VRL (Vila Real/ Régua/ Lamego) is a major strategic urban pole for sustaining population in rural interior areas. There is a regional plan approved since 1991 that for the first time assumed the need to develop these three towns as an important anchor in the interior Northern Portugal.

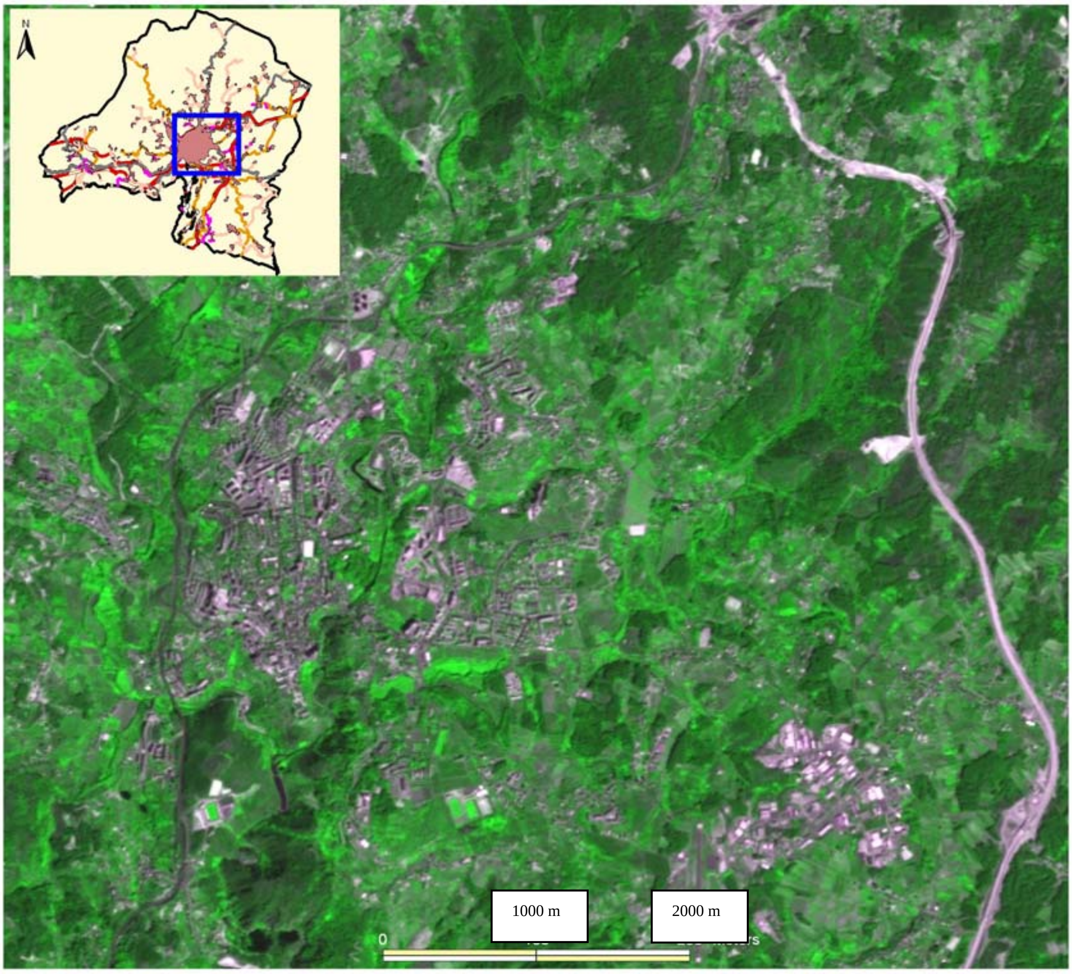


Figure 4: A sample of the satellite image (enlarged area)

CLUSTERING TRIALS

A set of five, ten, fifteen and twenty clusters were automatically processed within IDRISI program, according to an unsupervised classification based on the information in the colour composite image. An example is shown in Figure 6 for the maximum number of clusters. The contrast with a false colour image portrayed in Figure 5 highly denotes its accuracy.

The clusters are ranked in terms of how much the image they describe, using a histogram peak technique where a peak is defined as a value with a greater frequency than its neighbours on either side.

Once the peaks have been identified, all possible values are assigned to the nearest peak and the divisions between classes fall at the midpoints between peaks. Here a three-dimensional histogram is used because the composite is derived from three bands. An example is shown for the ten clusters identification where relationships between classes are automatically detected (see Figure 7).

The accuracy of the classification is not yet satisfactory with this previous supervised classification method (MAXLIKE). This is clearly visible when looking at first glance to figures 8 and 9.

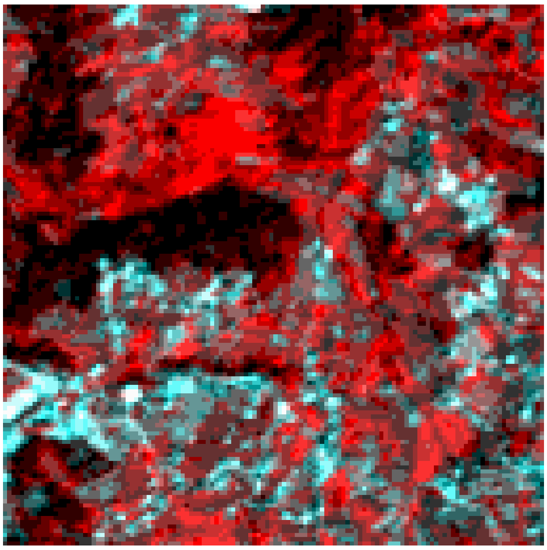


Figure 5: False Colour Image

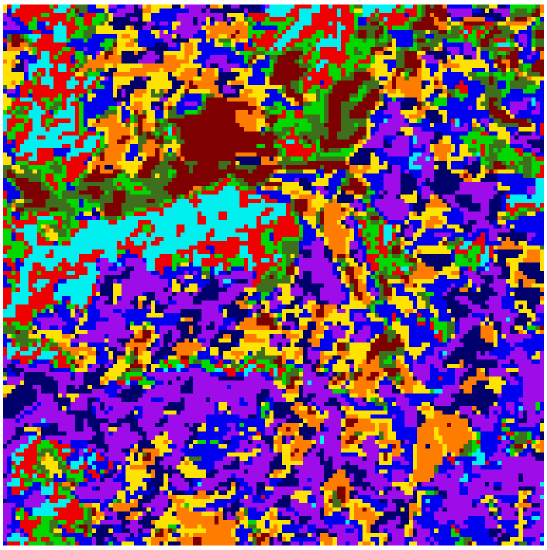


Figure 6: Cluster with 10 classes

```
Histogram of vreal_cluster10

Class   Lower Limit   Upper Limit   Frequency  Prop.    Cum. Freq.    Cum. Prop.
-----   -----------   -----------   ---------  ------   ----------   ----------

0         1.0000        1.9999        2602   0.1588        2602        0.1588
1         2.0000        2.9999        1896   0.1157        4498        0.2745
2         3.0000        3.9999        1396   0.0852        5894        0.3597
3         4.0000        4.9999        1114   0.0680        7008        0.4277
4         5.0000        5.9999        2962   0.1808        9970        0.6085
5         6.0000        6.9999        1080   0.0659       11050        0.6744
6         7.0000        7.9999        1091   0.0666       12141        0.7410
7         8.0000        8.9999        1349   0.0823       13490        0.8234
8         9.0000        9.9999        1804   0.1101       15294        0.9335
9        10.0000       10.9999        1090   0.0665       16384        1.0000


Class width       =        1.0000         Actual maximum    =       10.0000
Display minimum   =        1.0000         Mean              =        4.9984
Display maximum   =       10.0000         Stand. Deviation  =        2.9138
Actual minimum    =        1.0000         df                =         16383
```

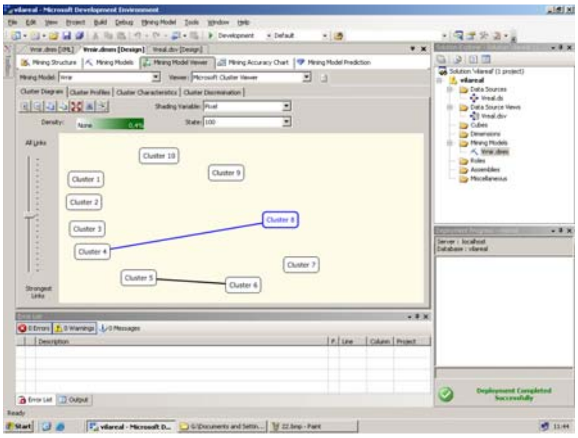


Figure 7: Results on clustering process – an example

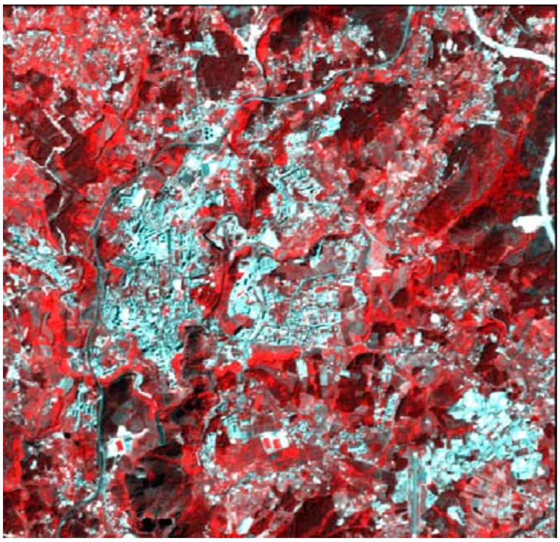


Figure 8: SPOT partial Image - False colour

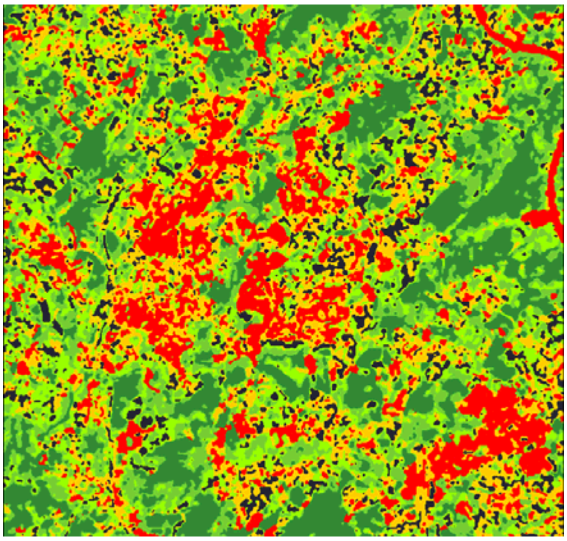


Figure 9: Classified partial Image (MAXLIKE)

A first numerical result shows the dominance of agricultural areas (37% of the total) in the sample image followed by vegetation, removed earth and urban areas, each class occupying similar areas, between 19,5% and 16% of the total.

The area occupied by pine plantation is the smallest of all with only 9% of the total area.

Table 3: Data on classified areas

Class	Count	%	Sq Km
Vegetation	71,865	19,45	7,19
Urban	59,076	15,99	5,91
Agriculture	137,375	37,18	13,74
Removed earth	67,860	18,36	6,79
Pine plantations	33,263	9,00	3,33
TOTAL	369,439	100,00	36,94

APLICATION OF THE DATA MINING METHOD

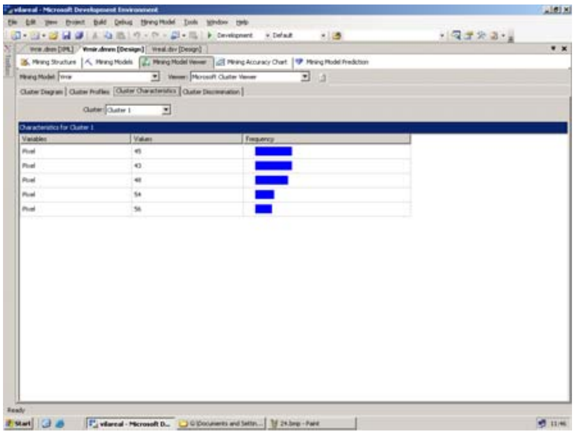
The *input data* is directly obtained from an Image Conversion tool (belonging to the Image Analyst framework) and produces an ASCII file with one column named Pixel with enough information to be used within a Data Mining tool.

In order to use the selected data mining tool, these data was further processed to import it to a data base (MS Access).

The overall process can be conducted by the Microsoft Business Intelligence Development Studio 2005 package which is an integrated package for data mining supporting a well known process model SEMMA - Selection, Exploration, Modification, Modelling, Assessment [11].

The data source used for the presentation of these histograms was the original (around 468854 registers) corresponding to one Satellite SPOT 5 image from September 2003. The filtered set contains 468854 registers. These registers were previously analysed. This section deals with clustering and histograms related to the initial data set corresponding to the image are presented.

Some histograms (see Figures 10 to 13) show better defined land-use classes than others. But very near albeit different reflectances can be of the same use. An example is agricultural use where some fields have been watered, thus humidity gives a different albeit very near reflectance from the fields with same crop but either wet or dry land. The same applies to roofs with different solar exposition, for example.



Mining Model: Vrnir

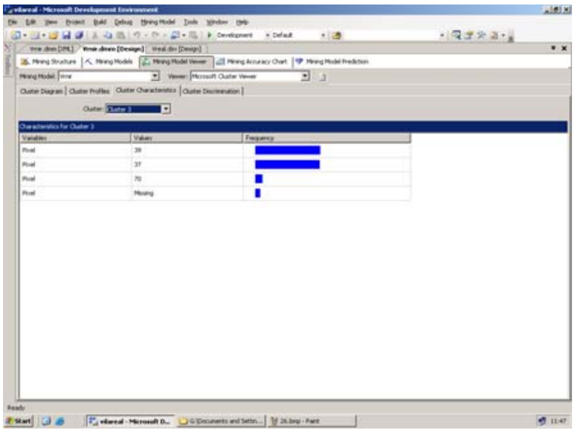
Viewer: Microsoft Cluster Viewer

Cluster: Cluster 1

Characteristics for Cluster 1

Variables	Values	Frequency
Pixel	45	▬
Pixel	43	▬
Pixel	48	▬
Pixel	54	▬
Pixel	56	▬

Figure 10: Histogram of class 1



Mining Model: Vrnir

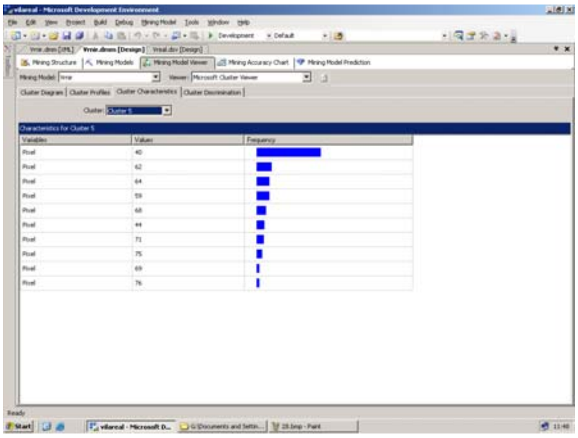
Viewer: Microsoft Cluster Viewer

Cluster: Cluster 3

Characteristics for Cluster 3

Variables	Values	Frequency
Pixel	39	▬
Pixel	37	▬
Pixel	70	▬
Pixel	Missing	▬

Figure 11: Histogram of class 3



Mining Model: Vrnir

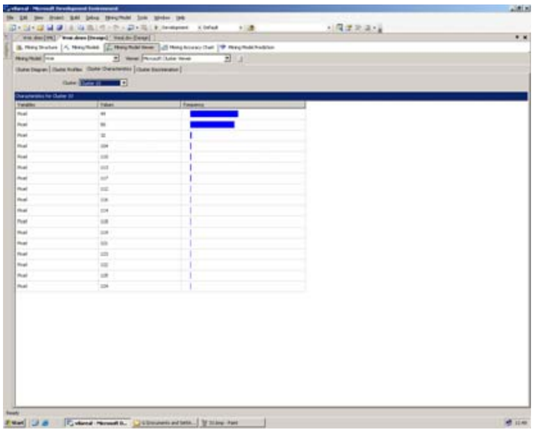
Viewer: Microsoft Cluster Viewer

Cluster: Cluster 5

Characteristics for Cluster 5

Variables	Values	Frequency
Pixel	40	▬
Pixel	62	▬
Pixel	64	▬
Pixel	59	▬
Pixel	68	▬
Pixel	44	▬
Pixel	71	▬
Pixel	75	▬
Pixel	69	▬
Pixel	76	▬

Figure 12: Histogram of class 5



Characteristics for Cluster 10

Variables	Values	Frequency
Pixel	44	▬
Pixel	86	▬
Pixel	32	▬
Pixel	104	▬
Pixel	110	▬
Pixel	113	▬
Pixel	107	▬
Pixel	102	▬
Pixel	116	▬
Pixel	114	▬
Pixel	118	▬
Pixel	119	▬
Pixel	[illegible]	▬
Pixel	123	▬
Pixel	122	▬
Pixel	128	▬
Pixel	124	▬

Figure 13: Histogram of class 8

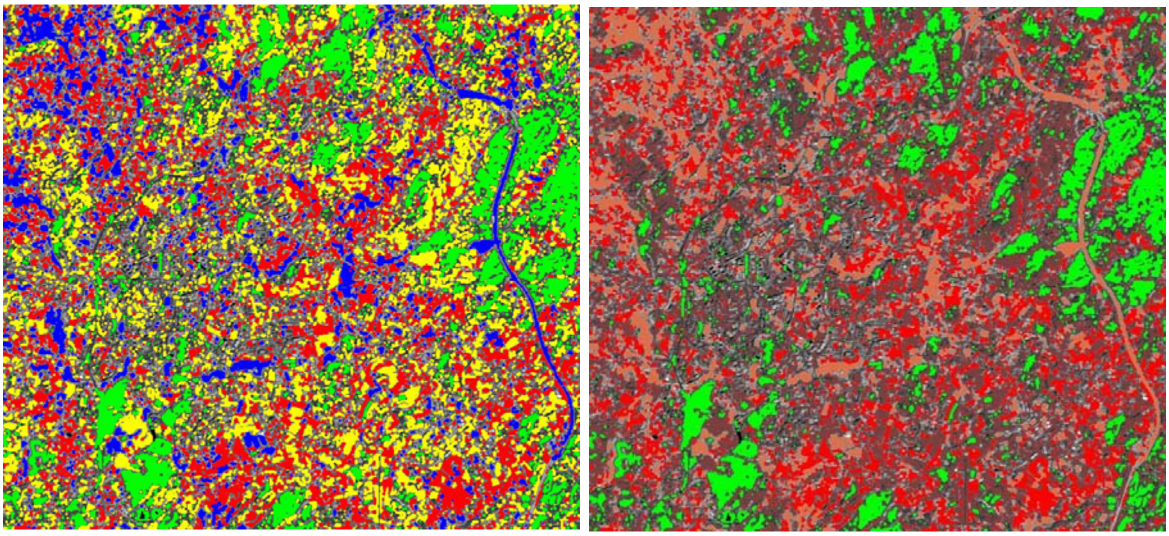
It was found that the segmentation result of each of the obtained classes has some associated noise. This is a typical situation [8,9,10]. Throughout a visual analysis, this noise corresponded to small areas, normally with a dimension lower than nine points or with short width lines. In order to avoid this problem, a morphological aperture filter was used. This filter which consists of a morphological erosion filter followed by a dilatation filter used a mask element 3x3, as shown below:

$$B = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

The overall operation of one image *A*, is thus defined as follows:

$$A \circ B = \left(A \Theta B \right) \oplus B$$

After that, the images were reclassified through neural networks processing where the entries are the average and standard deviation of near infrared image pixel value obtained for each one of the clusters areas previously defined (see Figures 14 a and b).



Figures 14a: Images resulting from Neural Networks classification from original data; 14b: with filtering process

CONCLUDING REMARKS

A first important outcome is then to have a good workflow for data mining and to have tested it against conventional methods.

From the analysis of applying this methodology to the satellite image SPOT 5 (2003) the following concluding remarks can be stated as major strengths:

- Despite the complexity of the technology involved, a user-friendly interface facilitates the implementation of the method;
- Support of several Data Mining techniques promotes efficient performance evaluation;
- The generalisation power of neural networks (through data generalization or inclusion of novel attributes) is very relevant for satellite images classification;
- "A priori" hypothesis on class distribution is not required;
- A small number of samples is enough to start the training process and to get first results;
- Neural networks produce more accurate results than conventional methods like MAXVER routine, with the same samples;
- With different hypothesis neural net perform better also.

The following weaknesses were also detected:

- the choice of the optimal neural network architecture is a task solved through trial and error;
- a bigger amount of samples, which improves the training process, require a very long response time, for this particular task.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the support of the Portuguese National Foundation for Science and Technology (FCT) under Project POCTI – CIVIL 49495, partially funded by ERDF.

REFERENCES

1. **Adriaans, Pieter and Zanting, Dolf**, Data Mining Using Enterprise Miner Software: A Case Study Approach, SAS Publishing. 2000.
2. **Cortez, Paulo**, Algoritmos genéticos e redes neuronais na previsão de séries temporais, Universidade do Minho, Braga, 1997.
3. **Cortez, Paulo and Neves, José**, Redes Neuronais Artificiais, Universidade do Minho, Braga, 2000.
4. **Costa, Paula**, Uma análise do consumo de energia em transportes nas cidades portuguesas utilizando Redes Neurais Artificiais, Universidade do Minho, Braga, 2003.
5. **Fayyad U. et al**, From Data Mining to Knowledge Discovery: An Overview. In Fayyad et al. (eds) Advances in Knowledge Discovery and Data Mining. AAAI Press / The MIT Press, Cambridge MA, 1996, pp 471-493.
6. **Filipe, V.**, Aplicação de Redes Neurais na análise de movimento, Universidade do Minho, Braga, 1997.
7. **Goebel M. and Gruenwald L.**, A Survey of Data Mining and Knowledge Discovery Software Tools. SIGKDD Explorations, v.1, nr.1, 1999, pp. 20-33.
8. **Gong, P.**, Integrated analysis of spatial data from multiple sources: using evidential reasoning and artificial neural network techniques from geological mapping. Photogrammetric Engineering & Remote Sensing, v. 62, nr.5, 1996, pp. 513-523.
9. **Logan, T. et al**, Artificial neural network classification using a minimal; training set: comparison to conventional supervised classification. Photogrammetric Engineering & Remote Sensing. v. 56, nr.4, 1997, pp. 1285-1294.
10. **Swain, P. H. & Ersoy, O. K.**, Neural network approaches versus statistical methods in classification of multiuse remote sensing data. IEE Transactions on Geoscience and Remote Sensing. v.28 nr.4, 1991, pp. 540-552.
11. http://www.sas.com/technologies/analytics/datamining/miner/index.html

AUTHORS INFORMATION

Júlia LOURENÇO
jloure@civil.uminho.pt
Universidade do Minho
CEC

Luís RAMOS
lramos@utad.pt
Universidade de Trás-os-Montes e Alto Douro
CETAV

Rui RAMOS
rramos@civil.uminho.pt
Universidade do Minho
CEC

Henrique SANTOS
hsantos@dsi.uminho.pt
Universidade do Minho
ALGORITMI

Delfim FERNANDES
delfimf@utad.pt
Universidade de Trás-os-Montes e Alto Douro
CETAV