International Opensource Geospatial Research Symposium 2009 (OGRS 2009)
Nantes, France, 8-10 July

An alternative approach to object oriented classification of historical black & white aerial photographs using GRASS GIS and R

**Eduardo Corbelle Rico**   **Rafael Crecente Maseda**

Research Group 1934 - Land Laboratory (LaboraTe)
University of Santiago de Compostela (Spain)
http://laborate.usc.es

LABORATORIO DO TERRITORIO (LaboraTe)

Some basic definitions...

Classification may be described as <u>the transformation of an image into a thematic map</u>.

Typically, each pixel in the image is assigned a *unique* class (**hard** classification), although other approaches allow for several labels being assigned (**soft** / fuzzy).

The kind of images more commonly subjected to classification usually present high **spectral resolution** (multi / hyperspectral). As spectral resolution increases, the ability to discriminate between different classes (separability) is also expected to increase.
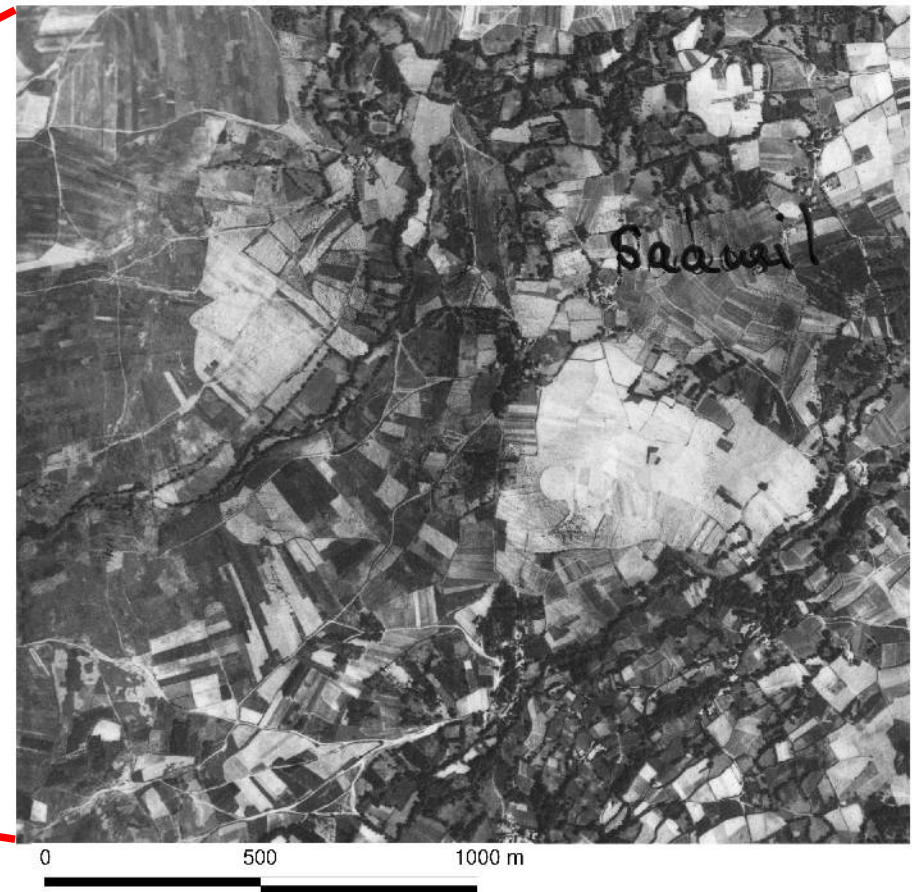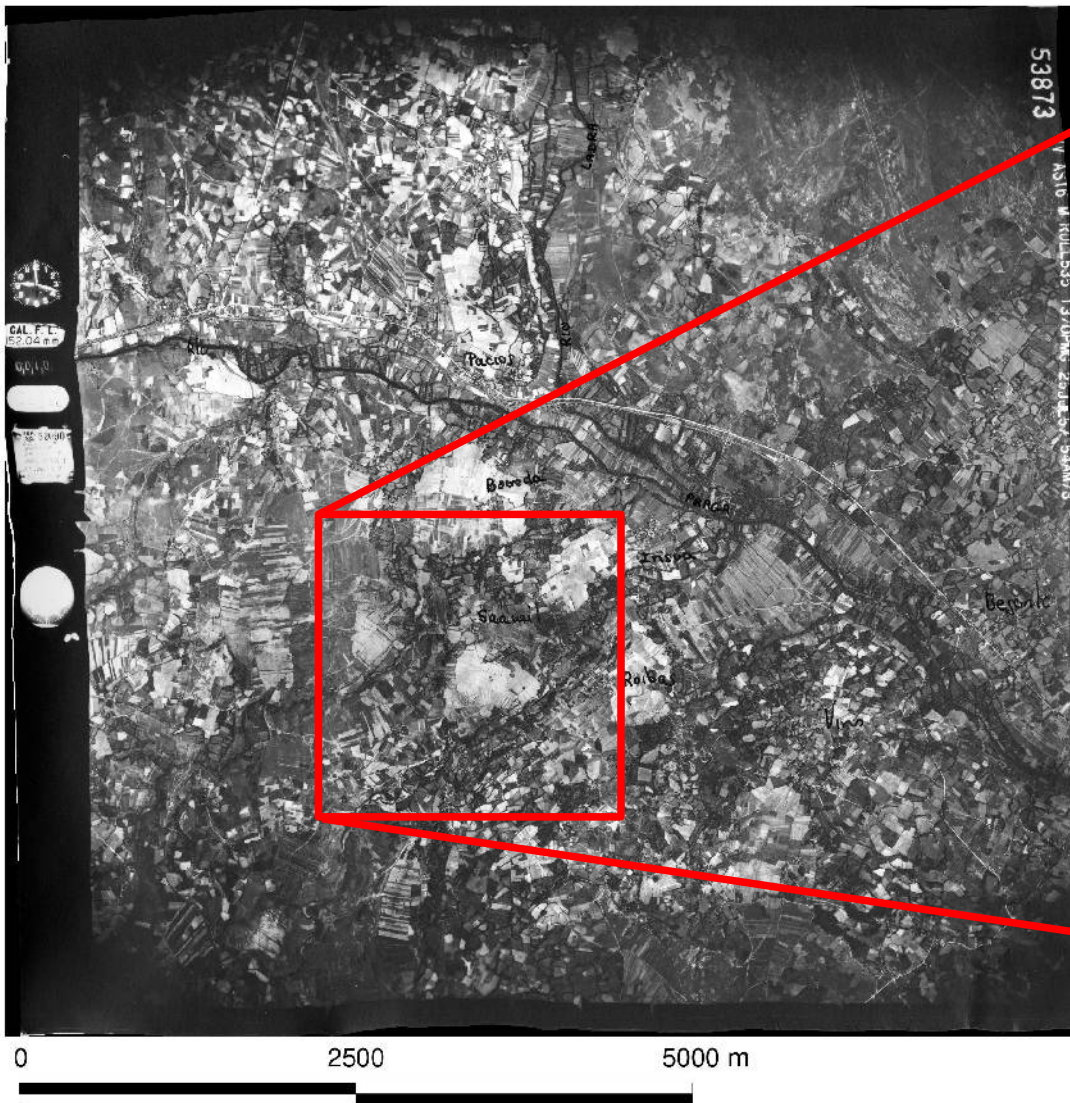
Example of separability of classes in two dimensions

high separability   low separability
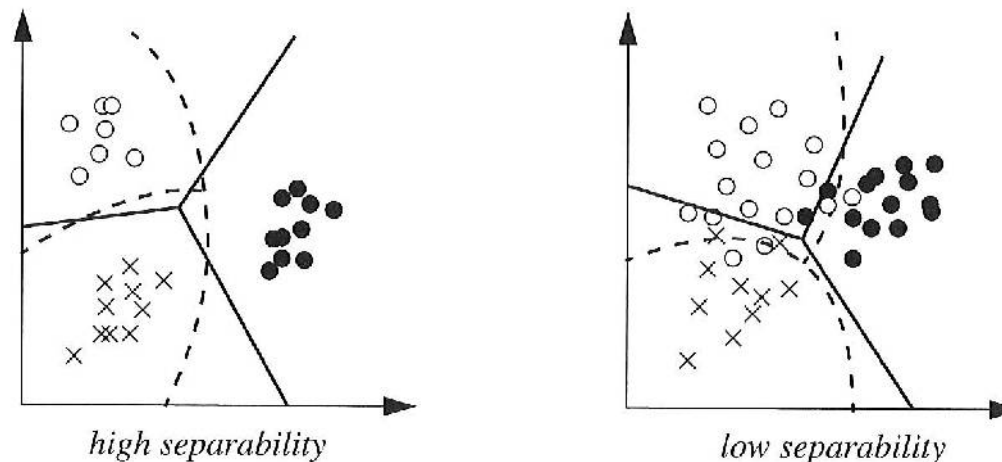
(taken from Schowengerdt 2007)

LABORATORIO DO TERRITORIO (LaboraTe)

The problem...

Land use / land cover change (LUCC) studies often depend on the use of black and white (panchromatic) aerial photographs. Separability is reduced to differences in grey levels.

Lack of adequate separability between classes causes "traditional" ***per-pixel*** schemes to show poor accuracy. In that context, neither parametric (e.g. nearest mean, maximum likelihood) nor non-parametric methods (e.g. *k nearest neighbours*) would probably offer good results. At least not much better than simple thresholding...
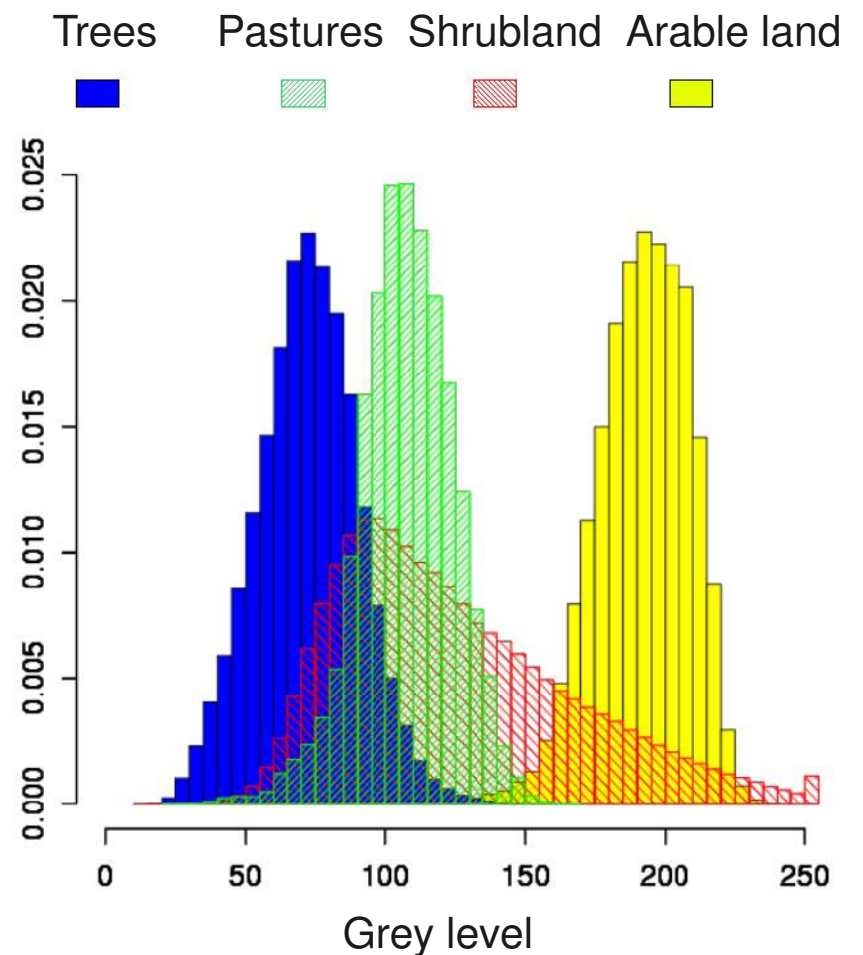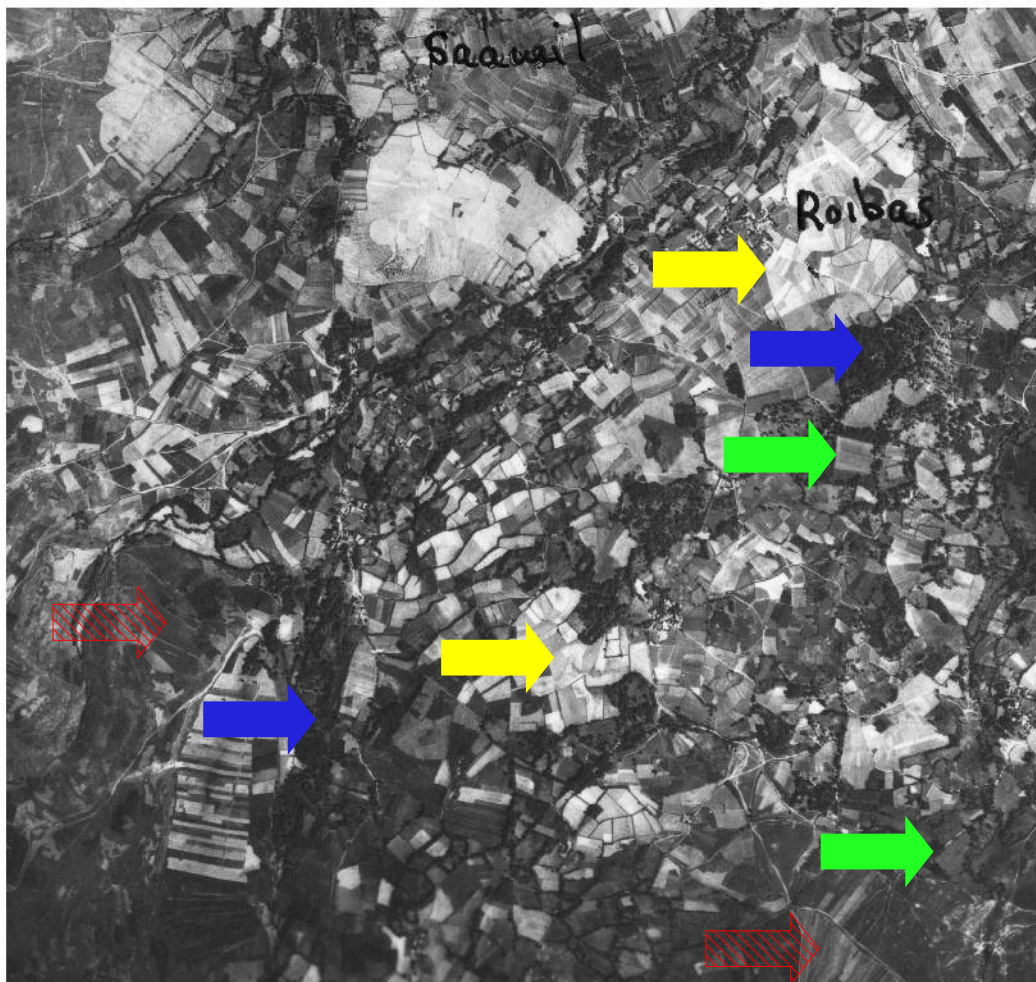
Object oriented classification...

In high spatial resolution imagery, real world objects are usually portrayed in the image as a **group** of statistically correlated (and spatially neighbouring) pixels.

Object oriented classification methods attempt to classify groups of pixels rather than individual ones, hoping to better represent the reality on the terrain.

Such methods often include a previous spatial-spectral **segmentation** process intended to divide the original image in a number of objects, which are then classified individually.

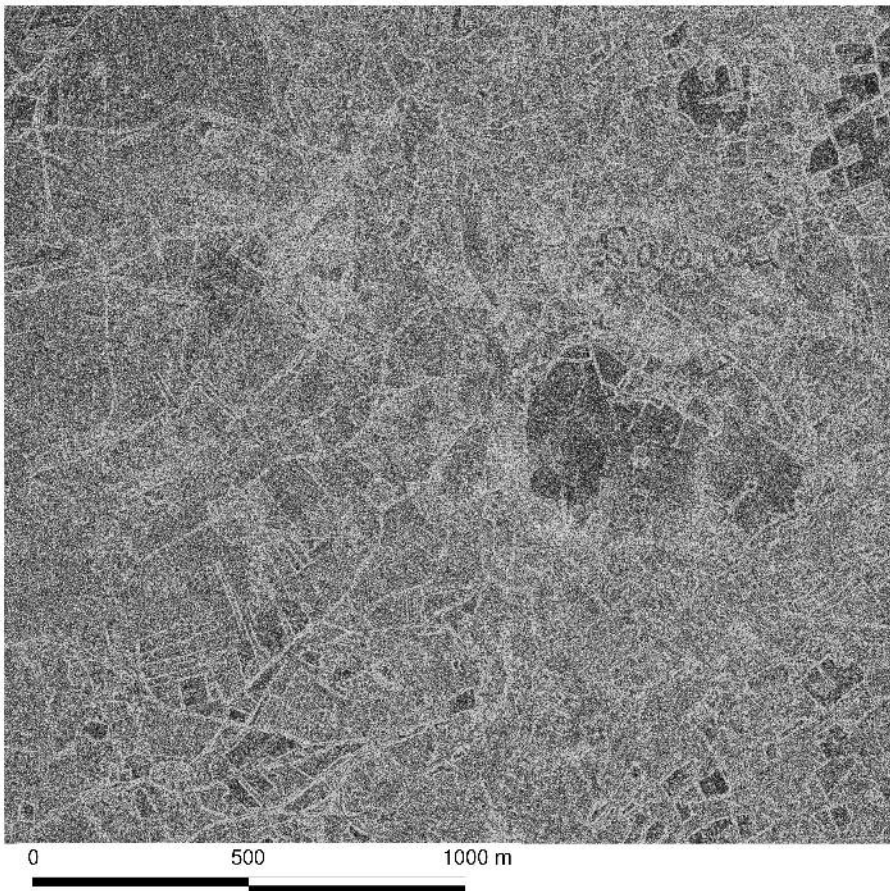Many algorithms exist, one is available in GRASS:

Sequential Maximum A Posteriori classification (SMAP)
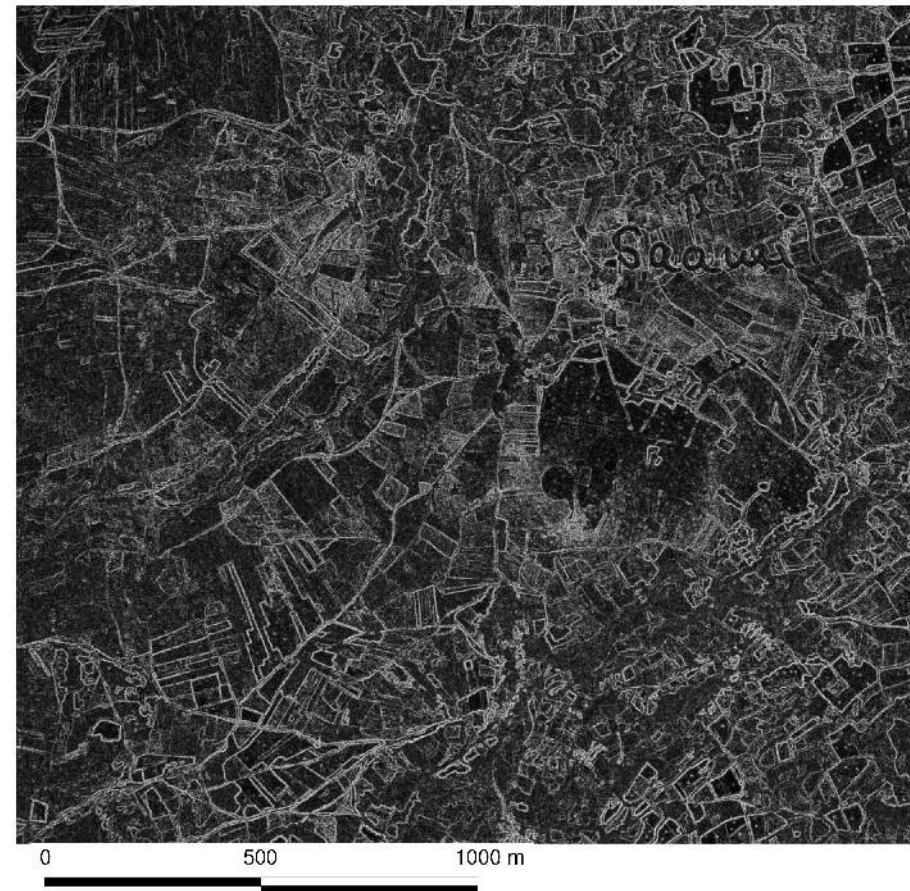
Textural information...

Given the low amount of information contained in the panchromatic band, the use of textural measures may be critical for the performance of classification.

GRASS (r.texture) offers up to 14 different measures that can be combined with different window sizes (e.g. sum entropy).

window size = 3   window size = 7

0   500   1000 m

0   500   1000 m

A window size of 5 was believed to be a good compromise: a smaller window showed no different values for the different covers, while larger ones increased the presence of borders (an undesirable effect) and also the processing time.

Sum entropy was selected among the available measures with the aid of PCA (i.pca)
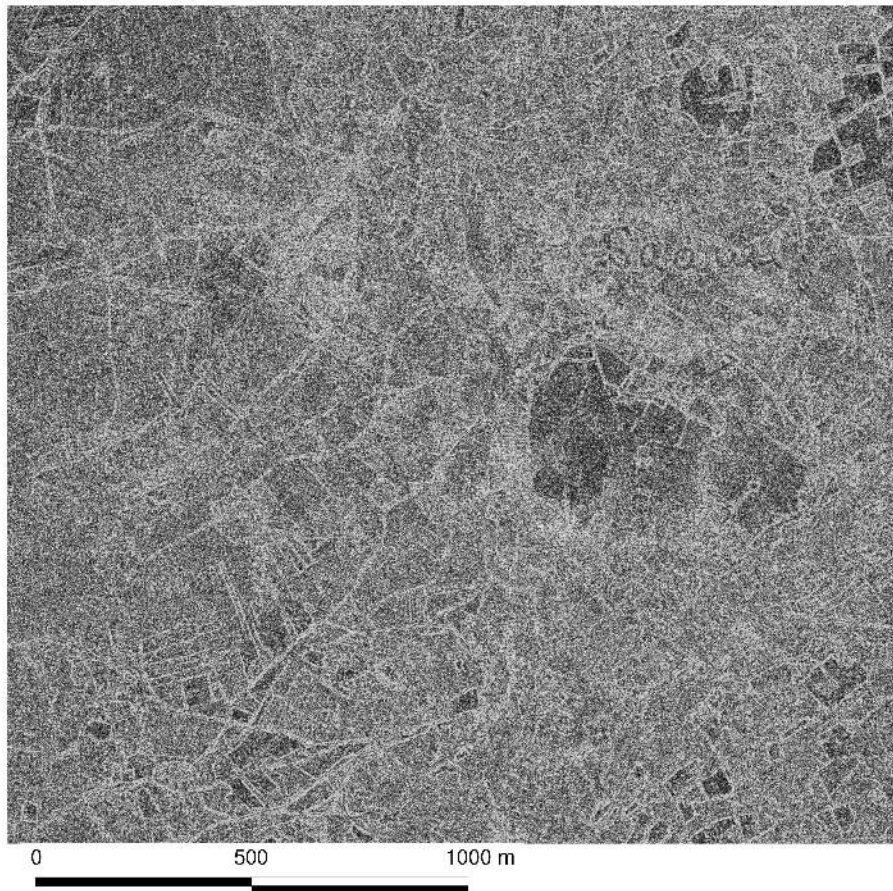
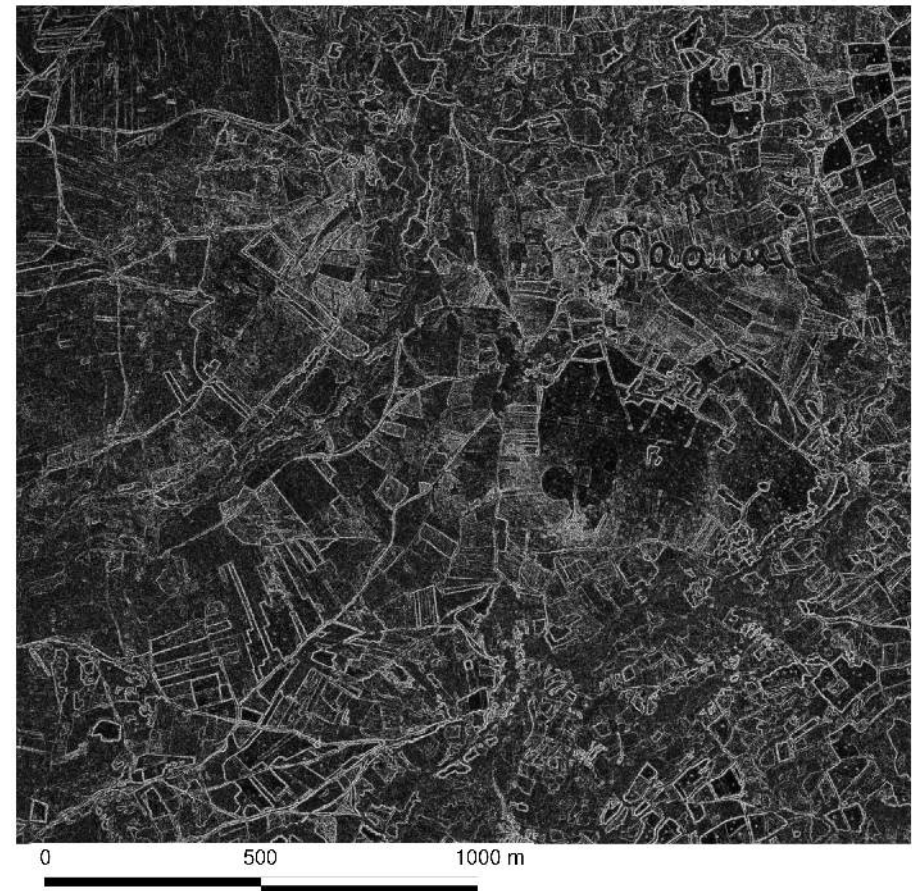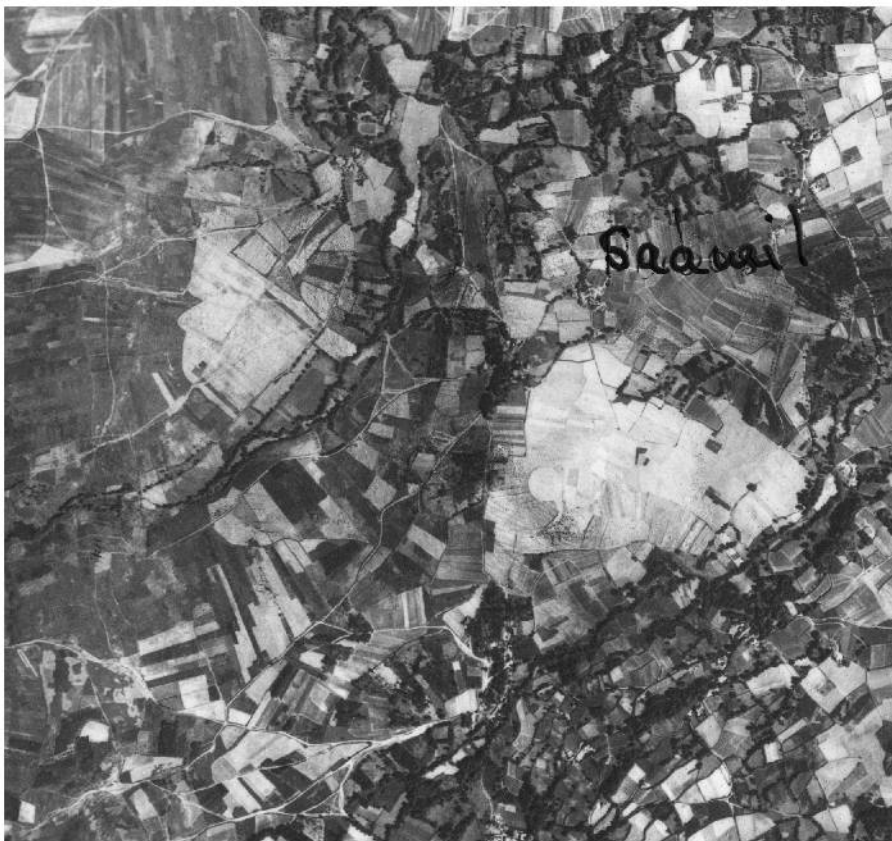window size = 3

window size = 7

A window size of 5 was believed to be a good compromise: a smaller window showed no different values for the different covers, while larger ones increased the presence of borders (an undesirable effect) and also the processing time.

Sum entropy was selected among the available measures with the aid of PCA (i.pca)

Original image

Sum entropy (window size = 5)
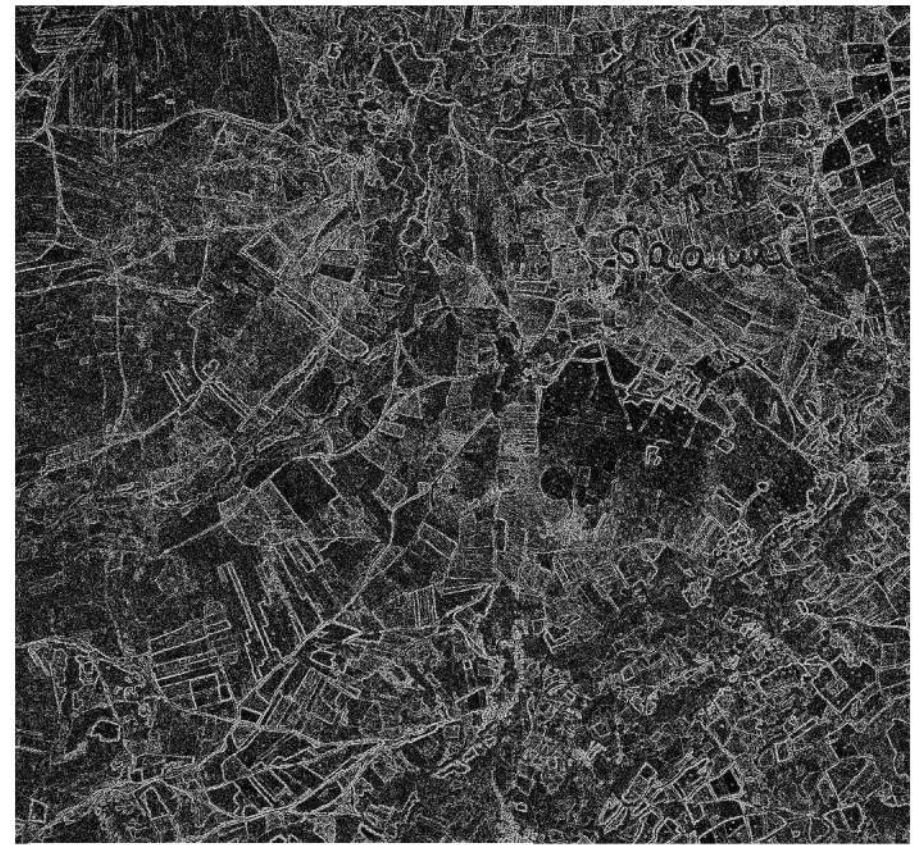
Limitations of SMAP...

Object oriented classification is usually divided into two **phases**:
>   (1) **segmentation** (delineation) of objects, and
>   (2) **classification**.

Many algorithms available in proprietary software carry out both phases separately, which allows for different data being used for each one.

SMAP offers no option but to carry out both segmentation and classification phases *simultaneously* and on the same data.

The use of **textural information** in the segmentation phase usually leads to wrong segments being created. This is particularly true for the **borders** between two different covers (textural measures show high values in these cases and may be confused with actual covers with high internal variability).

An alternative approach...

A suitable detour for this problem is the use of previously defined **vector polygons** instead of purposely-created segments. This possibility is included by default in some proprietary software packages.

Although not specifically programmed in GRASS, the use of previously defined vector polygons can be accomplished in R, e.g. using **supervised decision trees**.

SIGPAC (*Sistema de Información Geográfica de Parcelas Agrícolas*) is the Spanish **Land Parcel Information System** (LPIS), created to comply with EU regulations.

The basic cartographic unit is the cadastral **parcel**, but a smaller subdivision (**"recinto"**) was created to account for different uses within the same parcel.

We propose the use of "recintos" (vector polygons) as segments for the classification of historic land use, based on the assumption that the **structure of parcels** has not changed significantly in the period.

SIGPAC "recintos" on the original image...

LABORATORIO DO TERRITORIO (LaboraTe)
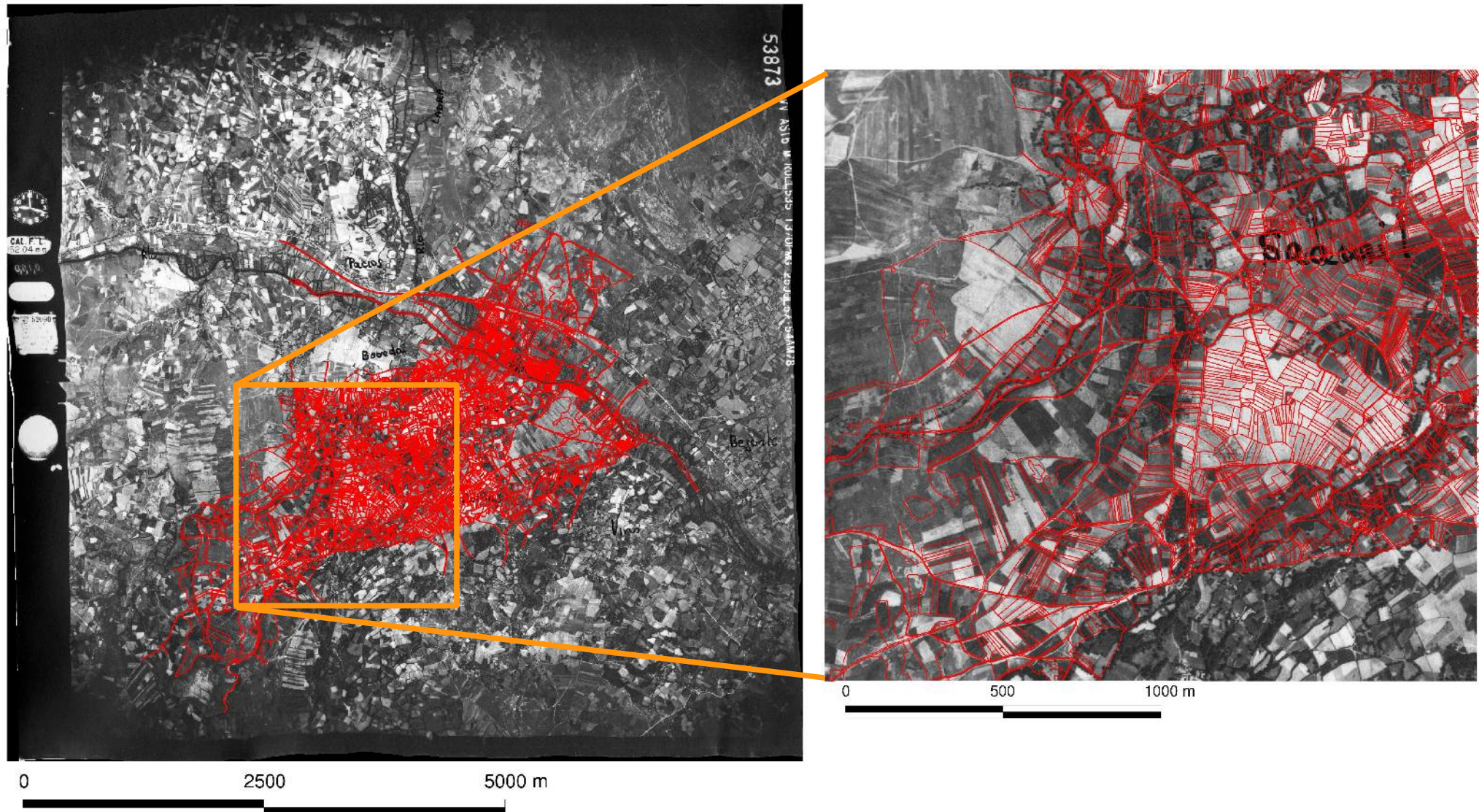
In summary...

We compare the results of **two classification approaches** performed on the group formed by the original image and the textural image:

- Standard use of SMAP

- Classification using SIGPAC vector polygons and supervised decision trees available in the R package RWeka (Hornik et al. 2009)

LABORATORIO DO TERRITORIO (LaboraTe)

Classification process (1): SMAP

a) Create an image group with the original image and the textural image(s) of choice (we used a PCA analysis with i.pca to select, among the different measures offered by r.texture, the one carrying more information).

b) Create a raster map of training areas (r.digit) for the 4 classes: (1) arable land, (2) pastures, (3) shrubland, (4) trees.

Classification process (1): SMAP

a) Create an image group with the original image and the textural image(s) of choice (we used a PCA analysis with i.pca to select, among the different measures offered by r.texture, the one carrying more information).

b) Create a raster map of training areas (r.digit) for the 4 classes: (1) arable land, (2) pastures, (3) shrubland, (4) trees.

c) Generate signature file from training areas (i.gensigset)

d) Call the classification algorithm (i.smap)

LABORATORIO DO TERRITORIO (LaboraTe)

Classification process (1): SMAP

a) Create an image group with the original image and the textural image(s) of choice (we used a PCA analysis with i.pca to select, among the different measures offered by r.texture, the one carrying more information).

b) Create a raster map of training areas (r.digit) for the 4 classes: (1) arable land, (2) pastures, (3) shrubland, (4) trees.

c) Generate signature file from training areas (i.gensigset)

d) Call the classification algorithm (i.smap)

e) For quality assessment purposes: create a vector set of 100 points randomly distributed over the area (v.random), and update its database with the manual interpretation of covers and with the results of classification (v.what.rast).

f) Load the database in R, either using readVECT6 (available in the spgrass6 package by Bivand 2008) or with read.dbf and cross-tabulate.
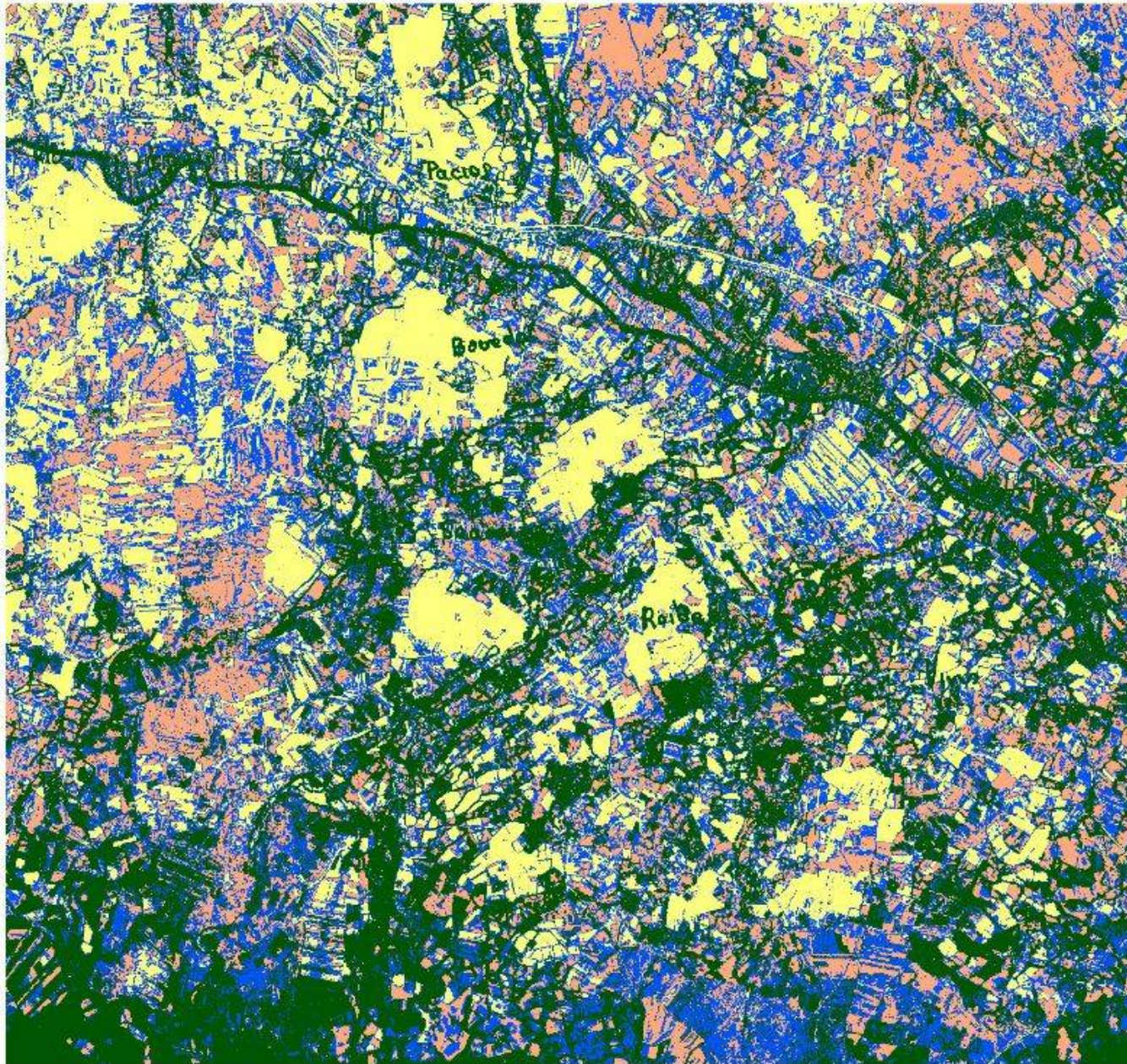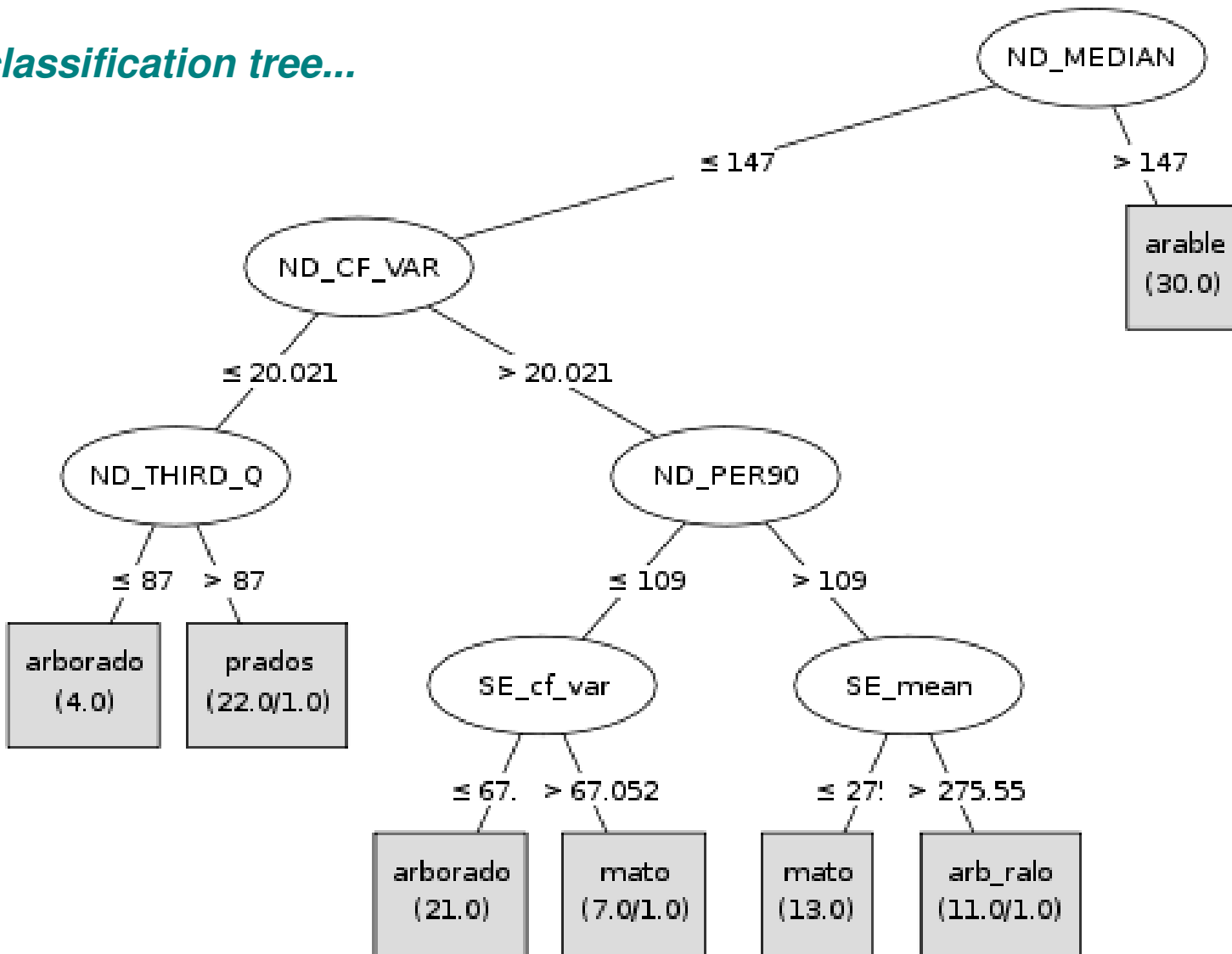
Classification process (2): SIGPAC polygons and supervised decision trees

a) Update the vector database with the values of the original image and the textural image(s) of choice (v.rast.stats).

b) Select a sample of training polygons and update the database with the class they belong to: arable land, pastures, shrubland, trees *PLUS* pastures with trees.

Classification process (2): SIGPAC polygons and supervised decision trees

a) Update the vector database with the values of the original image and the textural image(s) of choice (v.rast.stats).

b) Select a sample of training polygons and update the database with the class they belong to: arable land, pastures, shrubland, trees *PLUS* pastures with trees.

c) Load the database in R (readVECT6 *or* read.dbf)

d) Call the classification algorithm: a supervised decision tree is included in the base packaging of R (rpart), but we used J48 instead. J48 is the implementation of the -also free- C 4.5 algorithm (Quinlan 1993) in Weka and RWeka (Hornik et al. 2009, Witten & Frank 2005).

e) J48 allows for some degree of quality assessment using cross-validation (jackknife).

LABORATORIO DO TERRITORIO (LaboraTe)

The classification tree...

ND_... -> Grey levels
SE_... -> Sum entropy

LABORATORIO DO TERRITORIO (LaboraTe)
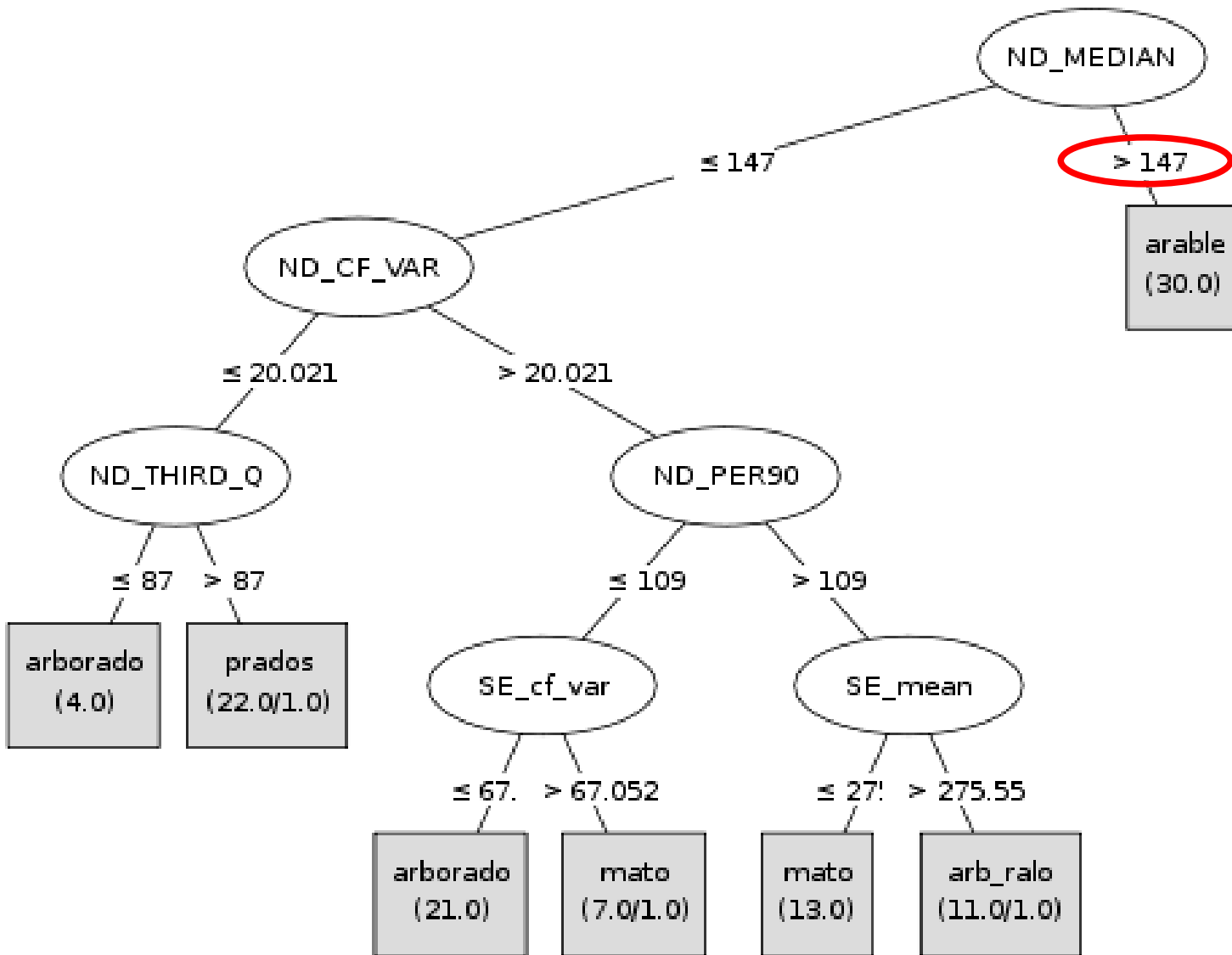
Arable land presents high reflectance

ND_... -> Grey levels
SE_... -> Sum entropy

LABORATORIO DO TERRITORIO (LaboraTe)

Arable land presents high reflectance

Pastures and some of the wooded parcels show low internal variability

Of the two possibilities, wooded parcels are generally darker
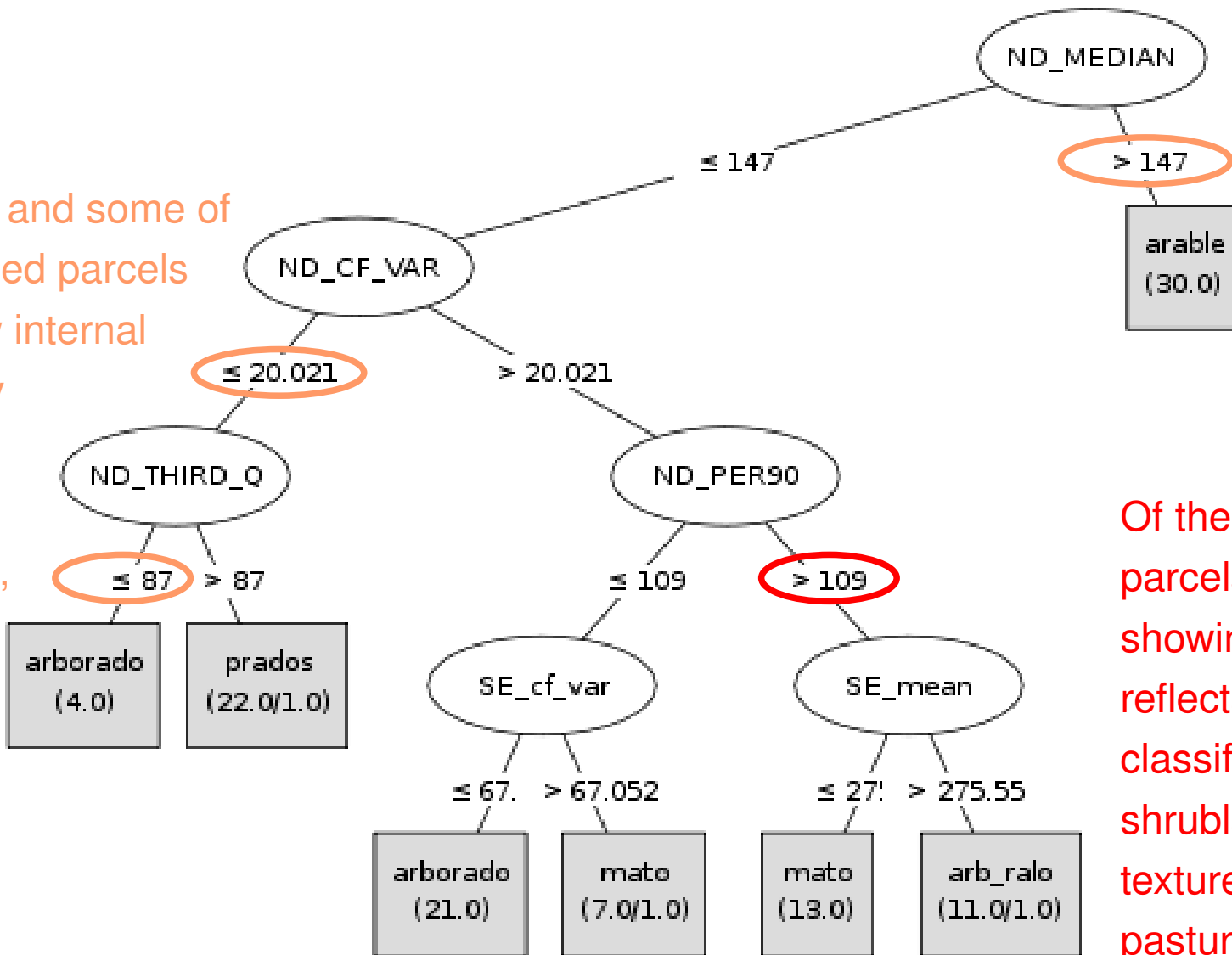
ND_... -> Grey levels
SE_... -> Sum entropy

LABORATORIO DO TERRITORIO (LaboraTe)

Arable land presents high reflectance

Pastures and some of the wooded parcels show low internal variability

Of the two possibilities, wooded parcels are generally darker

Of the remaining parcels, those showing higher reflectance values are classified either as shrubland (lower texture values) or pastures with trees

The darker parcels are classified as trees if they show more uniform values of texture or as shrubland otherwise
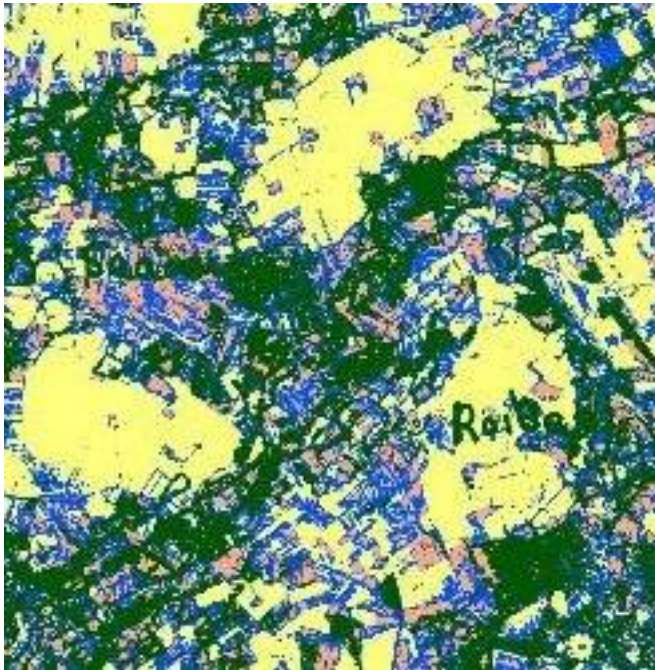
ND_... -> Grey levels
SE_... -> Sum entropy
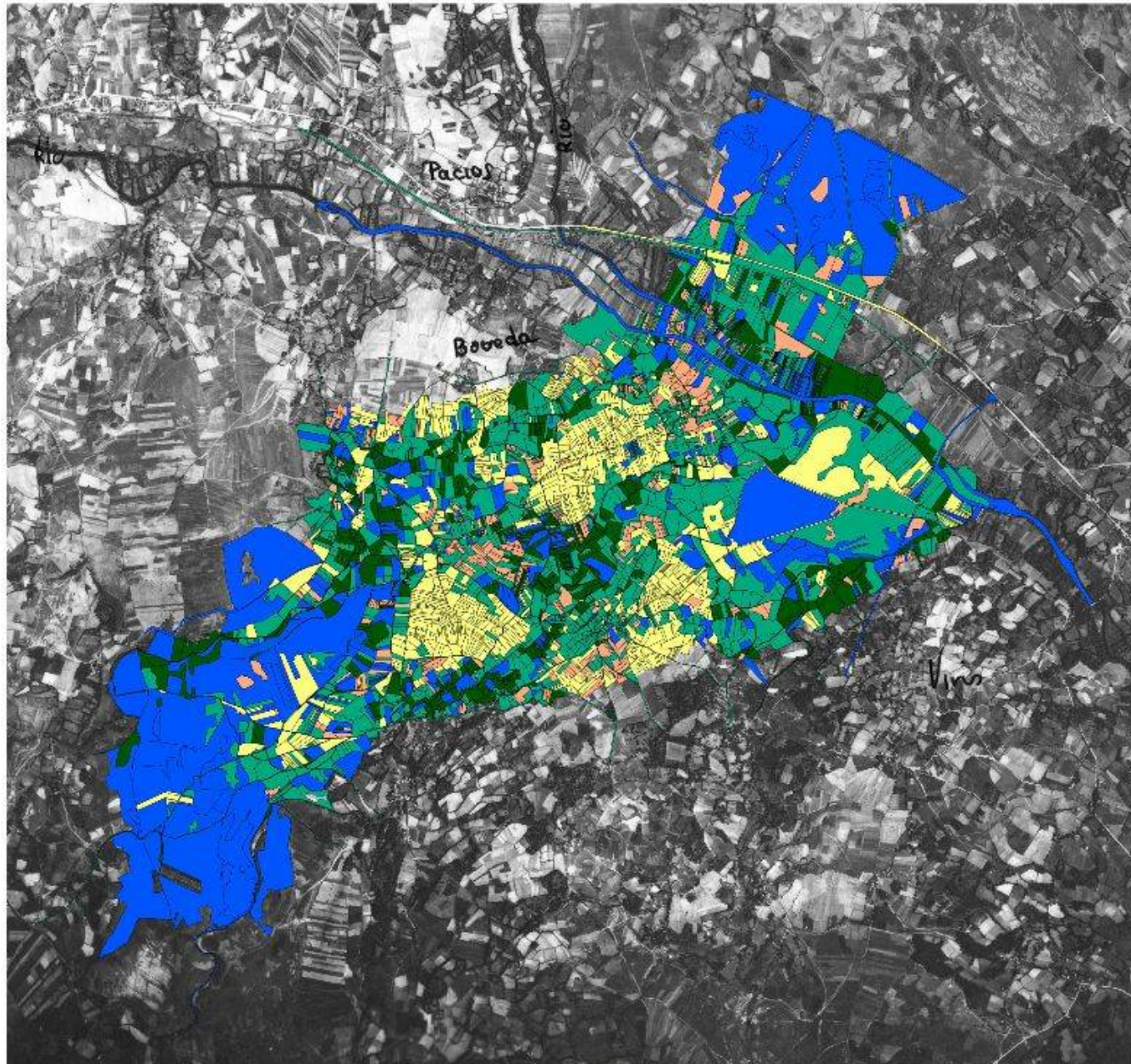
0 500 1000 m		Arable land	Shrubland	Pastures & Trees
		Pastures	Trees	

LABORATORIO DO TERRITORIO (LaboraTe)

SMAP

Confusion matrix

```
 a  b  c  d  e  <-- classified as
22  0     2  0 | a = arable
 1 13     2  1 | b = trees


 5  8    13  8 | d = shrubland
 2  3    12  3 | e = pastures
```
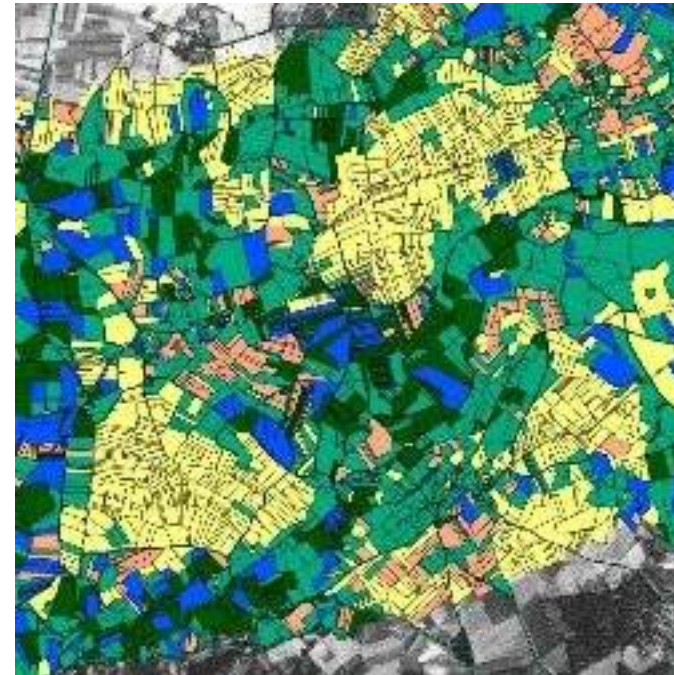
Global accuracy = 56/100 (56%)

(independent check points)

J48 and vector polygons

Confusion Matrix:

```
 a  b  c  d  e  <-- classified as
30  0  0  0  0 | a = arable
 0 23  1  2  1 | b = trees
 0  0 10  0  0 | c = past w/trees
 0  2  2 16  0 | d = shrubland
 1  2  0  2 16 | e = pastures
```
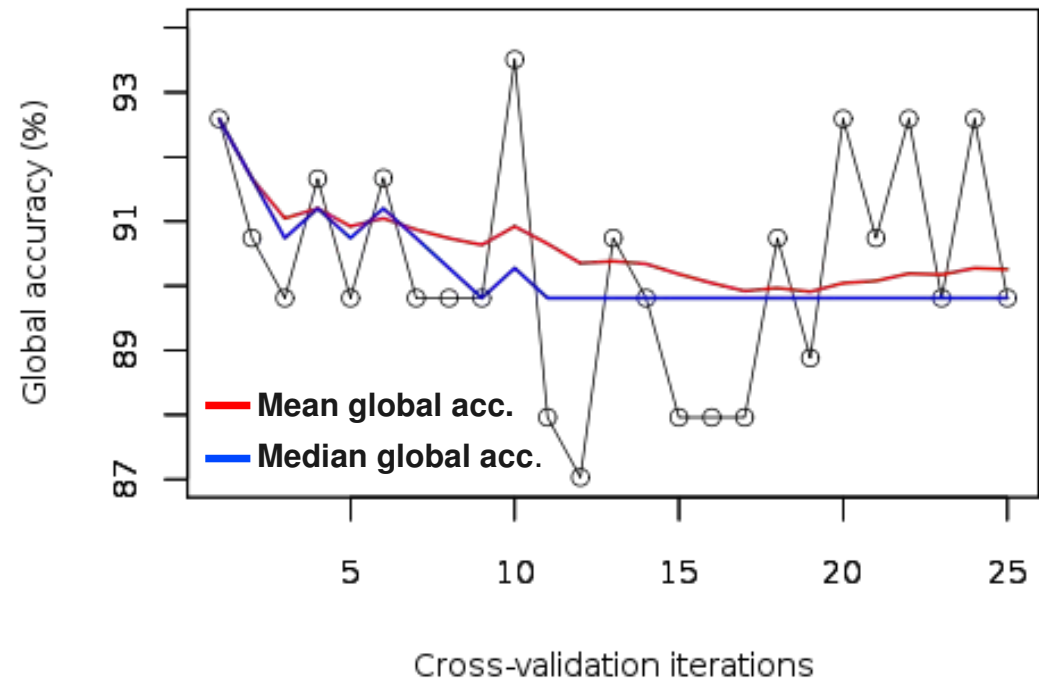
Global accuracy = 95/108 (88%)

(10 fold cross-validation)

Cross-validation iterations

Confusion matrix

```
 a  b  c  d  e  <-- classified as
22  0     2  0 | a = arable
 1 13     2  1 | b = trees


 5  8    13  8 | d = shrubland
 2  3    12  3 | e = pastures
```
Global accuracy = 56/100 (56%)

(independent check points)

Confusion Matrix:

```
 a  b  c  d  e  <-- classified as
30  0  0  0  0 | a = arable
 0 23  1  2  1 | b = trees
 0  0 10  0  0 | c = past w/trees
 0  2  2 16  0 | d = shrubland
 1  2  0  2 16 | e = pastures
```
Global accuracy = 95/108 (88%)

(10 fold cross-validation)

Some comments and conclusions...

The use of existing vector polygons to generate historic land use maps is a good option if the **structure of parcels** did not change significantly.

The use of previously defined polygons may be understood as a significant **simplification of the reality** portrayed in historic photographies, but it might still be a useful approach for LUCC studies.

Global **accuracy** of this alternative approach seems to be potentially higher than that attained with the default object oriented method available in GRASS (SMAP). Nevertheless, the issue of simplification already mentioned should be taken into account when comparing both figures. Also, the fact that one of them comes from an independent set of points while the other is generated by **cross-validation**, may pose problems for direct comparison.

Thank you for your attention...

e-mail: eduardo.corbelle@usc.es

References:

Bivand, R. (2008) *spgrass6: Interface between GRASS 6 geographical information system and R*. R package version 0.5-3. URL: http://grass.osgeo.org/

Hornik, K., Zeileis, A., Hothorn, T., Buchta, C. (2009) *RWeka: An R Interface to Weka*. R package version 0.3-15. URL: http://CRAN.R-project.org/

Neteler, M., Mitasova, H. (2004) *Open Source GIS: A GRASS GIS Approach (2nd ed.)*, Kluwer Academic Publishers, Boston, MA.

Quinlan, R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, San Mateo, CA.

Schowengerdt, R. A. (2007) *Remote Sensing: Models and Methods for Image Processing (3rd ed.)*, Academic Press, Burlington, MA.

Witten, I.H., Frank, E. (2005) *Data Mining: Practical Machine Learning Tools and Techniques. (2nd ed)*, Morgan Kaufmann, San Francisco.