Boosting : a classification method for remote sensing

Jean-Stéphane Bailly, M. Arnaud, C. Puech

Boosting: a classification method for remote sensing

Jean Stéphane Bailly* - Michel Arnaud - Christian Puech***

Abstract: This article sets out to demonstrate how boosting can serve as a supervised classification method, and to compare its results with those of conventional methods. The comparison begins with a theoretical example in which several criteria are varied: number of pixels per class, overlapping (or not) of radiometric values between classes, with and without spatial structuring of classes within the geographical space. The results are then compared with a real case study of land cover based on a multispectral SPOT image of the Sousson catchment area (South of France). It is seen that 1) maximum likelihood give better results than boosting when the radiometric values for each class are clearly separated. This advantage is lost as the number of pixels per class increases; 2) boosting is systematically better than maximum likelihood in the event of overlapping radiometric variable classes, whether or not there is a spatial structure.

** UMR3S CEMAGREF ENGREF Maison de Télédétection, 500 rue J. F Breton F-34093 Montpellier cedex 5*

{bailly, puech}@teledetection.fr

Tel: +33 (0)4 67 54 87 54 – Fax: +33 (0)4 67 54 87 00

*** CIRAD-TERA Campus International de Baillarguet TA 60/15 F-34398 Montpellier cedex 5*

michel.arnaud@cirad.fr

Tel: +33 (0)4 67 59 38 34 – Fax: +33 (0)4 67 59 38 38

1. Introduction

In many applications, it is often difficult to class, or more precisely to attribute an object (statistical unit) to a class (or group) defined a priori. In general, the term used for this is supervised learning.

This operation is conducted on a set of multivariate observations, and the methodology used is quite general: it consists in establishing a decision-making rule that fits the set of data as closely as possible.

Establishing these decision-making rules may involve a large number of techniques (Hastie *et al.* 2001) drawn from a wide range of fields such as statistics (discriminant factor analysis, decision trees, etc.) (Ripley 1996) or artificial intelligence (neural networks, etc.) (Bishop 1995, Ripley 1996).

Although these methods are very generally applicable and can be used in a wide range of sectors of application (economics, agronomy, sociology, geography, pedology, epidemiology, etc.), this article covers only applications directly linked to remote sensing, and we shall thus use the term supervised classification.

In view of the bulk of data to be analysed and their complexity, new methods have been developed in recent years in the data-mining sector and particularly in that of computer learning. This has recently become a highly active research sector (Dietterich 1999a) due to: 1) the meeting between researchers working in a range of fields (symbolic machine learning, calculator learning theory, neural networks, statistics, pattern recognition, etc.); 2) the application of learning techniques to new problems (the search for hidden knowledge in databases, language processes, robot control, combinational optimization, etc.); 3) the search for solutions to age-old problems (speech recognition, facial recognition, writing recognition, game theories, etc.).

This fruitful research has given rise to a certain number of supervised classification algorithms. We intend to present one that we feel to be well suited to the classification problems encountered in remote sensing: boosting. Boosting was first proposed by Freund and Schapire (1995, 1996), and is a procedure that results in a very precise law of prediction comprising a linear combination of relatively simple, moderately precise decision-making rules.

Although it is relatively recent, this tool has already been used for a certain number of geomatics applications: production of a land cover map under the MODIS project MOD12

Land Cover within IGBP[1], and work by Briem *et al.* (2001) on classifying multiple-sensor satellite images. Moreover, a study of boosting performance has been conducted, based on various configurations of two groups: diagonal linear border with and without intruders, alternate parallel strips and toruses (Arnaud *et al.* 2002).

It therefore seems particularly appropriate now to look closely at the tool's properties and assets, but also its limitations, in relation to applied work concerning remote sensing observations.

This article sets out to present the basic principles of the boosting algorithm and recall the characteristics of one of the algorithms most commonly used in remote sensing (MLC : maximum likelihood classification); to study how boosting works on well mastered theoretical examples and test its application to a real case (land cover in the Sousson catchment area); and, in both cases (theoretical and real), to compare the results obtained by boosting with those of conventional methods.

2. Boosting and conventional classification methods

2.1 *Problem, notation and algorithm*

Problem: the aim is to map land cover in a given zone, based on a remote sensing image. It is assumed that K classes of land cover have been defined for the zone. The image traditionally comprises a set of contiguous pixels making up the statistical units. Each pixel is characterized by its radiometric value, spatial position (geographical coordinates) and possibly by other information that may improve classification (soil type, crop history, toposequence, etc.). After field reconnaissance to link a pixel sample to one of the K classes,

[1] http://geography.bu.edu/landcover/userguidelc/lc.html and http://modis.gsfc.nasa.gov/sci_team/pubs/abstracts/MST-A0426.html

the aim is to establish a decision-making rule that attributes the non-sampled pixels to one of the classes as accurately as possible.

Basic, test and anonymous sets: the pixels are split into three sub-sets:

1. A set of **basic** pixels (also known as the training or learning set), for which the group number or class are known and with which the decision-making rule is **established** (classification);
2. A set of **test** pixels for which the group number is known, which are used to **test** the decision-making rule;
3. A set of **anonymous** pixels for which the group number is not known and which are intended to be **classed** in one of the K groups, using the decision-making rule established in (1).

Each of these three sets is necessary, particularly the test set, which enables cross-checking of the classification model. While the basic set is used to establish the decision-making rule, it is the test set that serves to validate it.

Notation: $\{(\mathbf{x_1}, y_1),\dots,(\mathbf{x_m}, y_m)\}$ is the basic set ($\{(\mathbf{x'_1}, y'_1),\dots,(\mathbf{x'_n}, y'_n)$ the test set respectively), where:

$\mathbf{x_i} = (x_i^1, x_i^2,\dots x_i^p) \in \mathbf{R}^p$. This is the vector of the so-called explanatory variables. The observed value (radiometry, geographical coordinates and other information of use for classification, both qualitative and quantitative) is designated x_i^j for pixel i and variable j designated x^j.

$y_i \in Y = \{1,\dots,K\}$; y_i is the variable to be explained. It corresponds to the number of the group to which pixel i belongs: K corresponds to the number of classes or groups.

Weak 'learner' and associated 'classifier':

Definition 1: a weak 'learner' is a function with real values:

$$h: \mathbf{R}^p \mathrm{x} Y \to \mathbf{R}$$

$$(\mathbf{x},\ell) \to h(\mathbf{x},\ell) \in \mathbf{R}$$

The values of $h(\mathbf{x},\ell)$ can be interpreted as follows:

- if $h(\mathbf{x},\ell)$ is positive, observation **x** is close to group ℓ. Conversely, it is far from group ℓ if $h(\mathbf{x},\ell)$ is negative;
- absolute value $|h(\mathbf{x},\ell)|$ characterizes the degree of confidence in the prediction: the higher the value, the 'safer' the prediction. It is not a probability but it has the same significance.

An example of weak 'learner' may come from decision trees (CART) (Breiman et al. 1984), corresponding to hyperlines in the space $\mathbf{R}^p$ of the variables. We expand further on that weak classifier example in remarks for table 1.

Definition 2: given $h_1,\ldots,h_T$, a set of T weak 'learners', it is function H with values within $Y = \{1,\ldots,K\}$that is the 'classifier' associated with weak 'learner' $\sum_{t=1}^{T} \alpha_t h_t$:

$$H : \mathbf{R}^p \to Y = \{1,\ldots,K\}$$

$\mathbf{x} \to H(\mathbf{x}) = \{\ell \in Y = \{1,\ldots,K\}$ such that $\sum_{t=1}^{T} \alpha_t h_t(\mathbf{x},\ell)$ reaches maximum}

Boosting algorithm: the initial algorithm AdaBoost (Freund and Schapire 1995, 1996), which only took account of two groups, was generalized through other algorithms, particularly AdaBoost.MH (Schapire and Singer 1999) with a view to addressing more general classification problems, particularly multi-class (more than two groups) and multi-label classification (possibility of assigning several groups to each observation simultaneously). It is this algorithm that we have used in this article. It is presented in detail in table 1 below.

[insert table 1 about here]

Some remarks concerning the algorithm (table 1):

- On each iteration t, AdaBoost.MH searches for a weak learner h_t from the observations in the basic set. Boosting is sufficiently flexible to handle any sort of weak learner, such as decision trees or neural networks. Many trials (Friedman *et al.* 1998) have shown that weak learners based on decision trees with a single level (CART) (Breiman *et al.* 1984) give good results. Algorithm AdaBoost.MH uses this method: at each step t, a function h_t is obtained that identifies both a variable x^j and a real value s_t^j, known as the threshold, relating to that variable. This value can be used to establish a distribution of the space $\mathbf{R}^p$ of the variables:

  $$\wp_0^t = \{\mathbf{x}=(x_1,\dots,x_p) \text{ such that } x^j < s_t^j \} \text{ and } \wp_1^t = \{\mathbf{x}=(x_1,\dots,x_p) \text{ such that } x^j \geq s_t^j \}$$

  The values assumed by weak learner h_t can be expressed as follows:

  $$\begin{array}{c|ccccc} & \text{Class } 1 & \dots & \text{Class } \ell & \dots & \text{Class K} \\ \hline \wp_0^t & c_{01} & \dots & c_{0\ell} & \dots & c_{0K} \\ \wp_1^t & c_{11} & \dots & c_{1\ell} & \dots & c_{1K} \end{array} \quad \text{where } h_t(\mathbf{x},\ell) = \begin{cases} c_{0\ell} & \text{if } \mathbf{x} \in \wp_0^t \\ c_{1\ell} & \text{if } \mathbf{x} \in \wp_1^t \end{cases}$$

- At the start of the process, all the basic observations have the same weight. At the end of each step, new weights are calculated in such a way that that of the wrongly classified items is greater than that of well classified items. At each step, these weights are taken into account by the weak learner (boosting-by-reweighting);

  The choice of coefficient α_t depends on the values taken by weak learner h_t. If it can take any real value, α_t is taken as 1. If h_t is made to take discrete values (-1 or +1), α_t is chosen so as to minimize normalization factor Z_t (Schapire and Singer 2000).

- *Advantages and limitations of AdaBoost.MH*: The algorithm has numerous advantages: it is quick in terms of calculation time and easy to program and use. The only parameter to set is T, the total number of iterations. T can be determined by doing an initial calculation ‘just to see’ and by looking at the percentage of well classified items for the basic and test samples. Subsequently, it is the T value for which the percentage of well classified items in the test sample no longer increases significantly that will be chosen. The weights calculated at each iteration are another boosting indicator: they offer a possibility of detecting irregular data (outliers), either due to an error in attributing them to a group or to the fact that the observation was very difficult to classify in one of the groups (Schapire 1999). On the other hand, AdaBoost is highly dependent on the basic sample and may not give satisfactory results if the training set contains an insufficient number of observations or if the data are very ‘noisy’ (Dietterich 1999b).

- *Testing and comparison with other algorithms*: *Adaboost* has been tested by numerous researchers in machine learning or artificial intelligence researchs (Dietterich 1999b, Drucker and Cortes 1996, Maclin and Opitz 1997, Quilan 1996) for several datasets (Blake and Merz 1998) and compared with learning methods such as C4.5 (Quinlan 1993). It has generally been the method that gives the best results. The performance of

AdaBoost.MH was compared with other methods for standard datasets in Yang's study (1999) and was one of the best.

2.2 *Using boosting*

Percentage of well classified items: there are two major indicators that can be used to analyse and interpret the scores proposed by boosting: the percentages of well classified items for the basic sample and the test sample. The test sample is essential, as it can be used to validate the results proposed by boosting on other observations not used to establish the classifier.

Determination of the value of T: the choice of criterion T is very important. Our tests showed that in most cases, the percentage of well classified items in the basic sample continued to rise, while the percentage of well classified items in the test sample, randomly selected, stabilized or even decreased. We therefore decided to give T the value that optimized the percentage of well classified items in the test sample.

While the percentages of well classified items in the basic sample and above all in the test sample are two good criteria for judging the relevance of the classification obtained by the final classifier, they are insufficient, since they are only global criteria, and do not provide any information on the observations themselves. As we shall see, boosting offers criteria that can be used to control the results at observation level.

Boosting weights: at each step t, the boosting algorithm calculates the values of weights D_t(i, ℓ) for each observation i in the basic sample in class ℓ. These values can be represented as follows:

$$D_t = \left(\begin{array}{c|ccc} & \text{Class 1} & \ldots & \text{Class K} \\ \hline 1 & D_t(1,1) & \ldots & D_t(1,K) \\ \vdots & \vdots & \ldots & \vdots \\ m & D_t(m,1) & \ldots & D_t(m,K) \end{array}\right) \quad \text{where } \sum_{i=1}^{m}\sum_{\ell=1}^{K} D_t(i,\ell) = 1$$

For each value of step t, it is possible to represent graphically the K*m values $D_t(i,\ell)$ that each represent the weight of each observation in the basic sample in class ℓ and to study any changes. In some cases, it will thus be possible to identify the observations that reveal high values that will necessitate a return to the data. High $D_t(i,\ell)$ weights will pinpoint the existence in the basic sample of one or even several outliers or wrongly classified data items.

Value of the learner: at step T (but also for any value t $\leq$ T), boosting provides the values of the weak learner $\sum_{t=1}^{T}\alpha_t h_t$. This function of $\mathbf{R}^{p}$xY and with values in $\mathbf{R}$ can be used to produce maps of each class ℓ=1…K from the columns in the matrix below. These maps will help in interpreting the attribution of each pixel (basic, test, anonymous or to be classified) in one of the K classes, taking account of both the sign of $\sum_{t=1}^{T}\alpha_t h_t$ (positive if i is close to class ℓ), but also of its absolute value (the higher the absolute value, the more accurate the prediction). For a given class ℓ, a distinction can be made between pixels with a strong chance of being classified in that classify (high absolute value and positive sign), pixels with a strong chance of not being classified in that classify (high absolute value and negative sign), and pixels for which there may be a degree of indecision (low absolute value).

$$h_T = \left(\begin{array}{c|ccc} & \text{Class 1} & \dots & \text{Class K} \\ \hline 1 & \sum_{t=1}^{T}\alpha_t h_t(1,1) & \dots & \sum_{t=1}^{T}\alpha_t h_t(1,K) \\ \vdots & \vdots & \dots & \vdots \\ m & \sum_{t=1}^{T}\alpha_t h_t(m,1) & \dots & \sum_{t=1}^{T}\alpha_t h_t(m,K) \end{array}\right)$$

Decision area: when only considering two variables in order to classify pixels (for instance the two radiometric variables XS2 and XS3), it may be worth visualizing the successive decision-making rules built from the weak learners obtained at each step t, within the variable area.

Importance of variables: at each step t, boosting selects one variable and one threshold. At the last step T, the contribution of each variable can be measured by counting the absolute and relative appearances of each one.

2.3. *The maximum likelihood method*

On remote sensing images, the most commonly used, standard supervised classification method is that of maximum likelihood. This is based on an assumption of the normality of the distribution of radiometric values for each class. In practice, this hypothesis is rarely tested, but the robustness of this method (deviation from the theoretical Gaussian model) means that it performs well.

We shall now briefly recall the principles of classification by maximum likelihood, which is widely described in the literature (Schowengerdt 1996, Monget 1997, Jia 1999). This method is based on calculating the probability of a pixel **x** being classified in class ℓ, $\ell = 1 \ldots K$. This probability uses Bayes' law:

$$P(\ell/\mathbf{x}) = \frac{P(\mathbf{x}/\ell).P(\ell)}{P(\mathbf{x})} \quad \text{avec} \quad P(\mathbf{x}) = \sum_{\ell=1}^{K} P(\mathbf{x}/\ell).P(\ell)$$

This shows that seeking probability $P(\ell/\mathbf{x})$ means maximizing the function $g(\ell,\mathbf{x}) = P(\mathbf{x}/\ell).P(\ell)$. This maximum is obtained by calculating the log-likelihood of $g(\ell, \mathbf{x})$:

$$\log(g(\ell,\mathbf{x})) = \log(P(\ell)) - \frac{p}{2}\log(2\pi) - \frac{1}{2}\log(|\hat{\Sigma}_\ell|) - \frac{1}{2}(\mathbf{x}-\hat{\mu}_\ell)'\hat{\Sigma}_\ell^{-1}(\mathbf{x}-\hat{\mu}_\ell)$$

In this case:

- $(\mathbf{x}-\hat{\mu}_\ell)'\hat{\Sigma}_i^{-1}(\mathbf{x}-\hat{\mu}_\ell)$ is the Mahalanobis distance from pixel **x** to classify ℓ;
- $\hat{\Sigma}_\ell$ is the experimental variance-covariance matrix of class ℓ in the radiometric variable area estimated from the basic sample;

- $\hat{\mu}_\ell$ is the experimental mean of class ℓ in the radiometric variable area estimated from the basic sample;
- $|\hat{\Sigma}_\ell|$ is the determinant of matrix $\hat{\Sigma}_\ell$;
- $P(\ell)$ is the prior probability of class ℓ in a Bayesian estimation. In this case, equiprobability prior of the classes is assumed. This hypothesis is open to debate in the application we intend to discuss, given that we already know the land cover pattern and the fact that temporal changes are only small on an overall level.

In the case of maximum likelihood, the decision-making rule will be: pixel **x** will be classified in class ℓ if it maximizes the quantity:

$$\log(g(\ell,\mathbf{x})) \text{ when } \ell \in \{1,\ldots,K\}$$

3. Applying boosting to a theoretical example

We tested the algorithm on several theoretical examples. For six land cover classes, we simulated the values of two variables, XS2 and XS3, combining 1) the more or less substantial sampling effort (N=12, 24 and 36 pixels per class), 2) whether or not the radiometric values overlapped, 3) whether or not geographical coordinates were used as additional explanatory variables, 4) in two clearly determined situations: one with a spatial structure and the other without. It is worth noting that sampling was interlocking: the small samples were part of the large ones. The results of these different types of configuration were compared with the conventional classification techniques found in standard software.

3.1 Without geographical coordinates:

We intend to present the results of the following two examples: the first will look at a situation in which the surrounding polygons of the six classes in the radiometric variable area do not intersect (figure 1). The second will look at the case of a non-vacant radiometric variable overlap between the classes. In both cases, we shall study the effect of increasing the number of pixels in the classes.

[Insert figure 1 about here]

3.1.1. *Without radiometric variable overlapping*:

In the six theoretical classes studied (forest, dense, medium and sparse vegetation, bare soil and water), shown in figure 1, we drew at random (with replacement) the whole XS2 and XS3 radiometric values of N = 12, 24 and 36 pixels respectively, in order to constitute a basic sample and a test sample. These samples were interlocked.

Percentage of well classified items: when the number N of pixels per class was equal to 12, the percentage of well classified items for the basic sample reached 100% at step 8, while the percentage of well classified items for the test sample reached a maximum value of 87.5% at step 16 (figure 2). When N was 24 and 36 respectively, the percentage of well classified items for the basic sample was 100% at steps 16 and 33, while the maximum percentage of well classified items for the test sample was 89.6 and 92.1% at steps 15 and 59. Within this range of variation in N, it is seen that when N increases, the percentage of well classified items also increases, but that this also takes a greater number of steps.

[Insert figure 2 about here]

Decision area: within the space formed by pair (XS2, XS3) it is possible to visualize at each step t the class partitioning established using weak hypotheses and based on the values of the basic sample. We visualized this partitioning for N = 12 at step t = 16 (i.e. when the percentage of well classified items for the test sample was maximum), and the pixels of the test sample (figure 3). This graph serves to identify the radiometric variable zones in which the test or anonymous pixels will be inappropriately classified. It reveals that this division closely fits the values in the basic sample, which is what makes the boosting method highly dependent on the basic data.

[Insert figure 3 about here]

Boosting weights: when N = 12, we represented the maximum and minimum weights of the pixels calculated at each step (less than or equal to 8) and in each class (figure 4). In the bare soil class ('*e*'), it is pixel 52 that stands out, with a high weight from t = 6 on. If one refers to its position within the radiometric variable area, it is seen to be on the far left of class '*e*' (figure 1, left) and is also very close to class 'a'. The algorithm will therefore give this pixel a high weight, resulting in a search for weak learners with a view to assigning pixel 52 to the right group (*i.e.* '*e*'). At step t = 8, although all the pixels in the basic sample scored highly, the absolute value of the weak learner for pixel 52 and for class '*e*' is still low. In subsequent steps, boosting will therefore attempt to increase this value to make the prediction more accurate. We observed that the pixel weights were relatively low in the other classes.

One of the main merits of boosting weights is to check the coherence or validity of the values of an observation in a group. This may therefore be a useful criterion in seeking out irregular data (outliers).

[Insert figure 4 about here]

3.1.2 *With radiometric variable overlapping*:

Percentage of well classified items: when there is a non-vacant intersection between the six classes and they overlap within the radiometric variable area (figure 5), the results of boosting deteriorate significantly: the percentages of well classified items in the basic sample decrease overall. In our case, they were 98.6% at step 86, 84.7% at step 95 and 76.4% at step 100 when the number of pixels was N = 12, 24 and 36 respectively. For the test sample, the percentages of well classified items fluctuated around 50% irrespective of the number of pixels per class: 50% at step 61 for N = 12, 50.7% at step 23 for N = 24 and 48.6% at step 28 for N = 32 pixels.

[Insert figure 5 about here]

In the event of significant overlapping between the classes and if geographical coordinates are not used, the boosting algorithm is apparently not capable of effectively classifying items.

In general, in addition to the radiometric values attached to each pixel, its position within a geographical area is always known. While conventional classification techniques such as maximum likelihood have difficulty taking account of this information due to the implicit statistical hypotheses assumed in the laws concerning the variables, boosting can take it into account more easily.

3.2 *With geographical coordinates and marked spatial structuring of the classes*:

We intend to look at how boosting performs in two very different geographical situations: that of a very strong spatial structure and that of an almost random distribution of classes. We introduce here geographical coordinates as demonstrative variables while the geographical space can be easily interpreted. We intend to assess with geographical coordinates how boosting is suitable for classification with specific or complex spatial structure features in the hyperspace of variables.

3.2.1. *Without radiometric variable overlapping*:

As we knew the radiometric values of the pixels for the two channels XS2 and XS3, we drew at random (but without replacement) the value of the coordinates of N = 12, 24 and 36 pixels for the basic and test samples within a geographical area with a very marked spatial structure. Figure 6 below shows the pixels of the basic sample for N = 12.

[Insert figure 6 about here]

Percentage of well classified items: for the basic sample, the percentage of well classified items was 100% at steps 8, 10 and 23 for N = 12, 24 and 36 pixels respectively per class. For the test sample, for N = 12, 24 and 36 respectively, the percentage of well classified items was 87.5% at step 46, 93.8% at step 16 and 97.2% at step 93.

Boosting weights: for N = 12, we represented the maximum and minimum weights of pixels in classes '*d*' and '*e*' for which the maximum value was high (figure 7). This identified two pixels: 52 and 46. Pixel 52, which is mentioned above, had a high weight in the bare soil class ('*e*'), while pixel 46 had a high weight in class '*d*' as well as in class '*e*'. Pixel 46 thus revealed another effect of boosting weights: they switched pixel 46 to class '*d*' from class '*e*'. It is to be noted that these two pixels are both on the border between these two classes, in both the radiometric variable and geographical areas (figures 1 and 6).

[Insert figure 7 about here]

Contribution of variables: if the radiometric values do not overlap, the geographical coordinates Xcoord and Ycoord have relatively little influence (around 20%) when N = 12 or 24, when establishing the decision-making rule constituted by boosting. Once N is greater (N = 36), the geographical coordinates become more important: their influence increases to around 37% (table 2)

[Insert table 2 about here]

In the event of a strong spatial structure and no overlapping between radiometric variable classes, using the geographical coordinates of pixels improves classification results, provided the number of pixels per class is not too small.

3.2.2. *With radiometric variable overlapping*:

Percentage of well classified items: in the event of a spatial structure, the results concerning the percentage of well classified items from the test sample are obviously less good with overlapping than without, but are still better than those obtained with overlapping but without using the geographical coordinates. For the basic sample in our case, the percentage of well classified items was 100% at steps 27 and 58 and 99.1% at step 88 for N = 12, 24 and 36 respectively. The percentages for the test sample were 69.4% at step 49, 81.9% at step 45 and 82.9% at step 30 for N = 12, 24 and 36 pixels per class respectively. The boosting algorithm thus compensates relatively well for overlapping radiometric values by using the geographical coordinates.

Importance of variables: in the event of overlapping radiometric values, the geographical coordinates are almost as important as the radiometric values (table 3).

[Insert table 3 about here]

In the event of a strong spatial structure and overlapping radiometric variable classes, using the geographical coordinates of the pixels improves the results of classification, irrespective of the number of pixels per class.

3.3 ***With geographical coordinates and without a spatial structure*:**

We drew at random (without replacement) the coordinates of N = 12, 24 and 36 pixels in the basic and test samples interlocked within a geographical area without a spatial structure or at least with a much more divided structure than before. This area was constituted by random distribution of 36 adjacent large plots representing each of the six land cover classes times six replicates. Figure 8 below shows the pixels of the basic sample for N = 12.

[Insert figure 8 about here]

3.3.1. *Without radiometric variable overlapping*:

Percentage of well classified items: the results were almost the same as when geographical information was not taken into account and the radiometric variable classes did not overlap (see above, 3.1.1). The percentages of well classified items for the basic sample were 100% at step 8, 10 and 23 for 12, 24 and 36 pixels per class respectively. For the test sample, for N = 12, 24 and 36 pixels respectively, the percentage of well classified items was 87.5% at step 16, 89.6% at step 57 and 93.5% at step 94.

Boosting weights: the results were again similar to those obtained without taking account of geographical information and in the event of non-overlapping radiometric variable classes (see above, 3.1.1).

Importance of variables: in the absence of overlapping between radiometric values, the geographical coordinates (Xcoord and Ycoord) had no effect when N = 12 and contributed for 14 and 27% when N = 24 and 36 respectively (table 4).

[Insert table 4 about here]

In the event of a lack of spatial structure and of overlapping between the radiometric variable classes, using the geographical coordinates of the pixels does not improve classification results.

3.3.2. *With radiometric variable overlapping*:

Percentage of well classified items: the results were better than with overlapping radiometric variable classes and without taking account of spatial information (see above, 3.1.2). The percentage of well classified items for the basic sample was 100% at steps 36 and 78 for N = 12 and 24 respectively and 97.7% at step 83 for N = 36. As regards the test sample, the percentages of well classified items were 62.5, 69.4 and 76.9% for N = 12, 24 and 36 respectively at steps 40, 66 and 98.

Importance of variables: in the event of overlapping radiometric values, the geographical coordinates (Xcoord and Ycoord) took over and intervened in the search for weak learners, in direct relation with N and with a contribution amounting to 45% when N = 12, 54.5% when N = 24 and 55.1% when N = 36 (table 5).

[Insert table 5 about here]

In the event of a lack of spatial structure and of overlapping radiometric variable classes, using the geographical coordinates of the pixels improves classification results. However, the improvement is much greater in the event of a spatial structure.

In the various situations mentioned above, maximum likelihood was used. Table 6 below gives the comparative results of the two methods. In terms of the percentage of well classified items in the test sample:

- on the whole, boosting generally gave much better results when spatial information (pixel coordinates) was used, except in the event of a simultaneous lack of spatial structure and of overlapping radiometric variable classes;
- maximum likelihood gave better results than boosting in the event of a clear separation of the radiometric values of each class. This advantage disappeared as the number of pixels in the class increased;
- boosting was systematically better than maximum likelihood when there was overlapping between the radiometric variable classes, irrespective of whether there was a spatial structure.

[Insert table 6 about here]

4. Application:

4.1 *Sousson catchment area landscape context*

We applied the method to the classification of land cover in the Sousson catchment area, as per a simple nomenclature similar to the theoretical case quoted above. The catchment area covers an area of 120 km², is in south-western France and has the elongated north-south shape characteristic of the region (see figure 9). Studies are under way in the catchment area on the transfer of pollutants of agricultural origin, and in particular require information on land cover at plot level. The site, for which exhaustive field information is available that tallies with the date of the processed image (18 August 1996), has already been used for comparisons of image classification methods (agricultural unit mode classification (Colin 2002), classification and prediction using temporal information (Dufour 2001)).

Land cover is structured within the geographical area by virtue of the simple geometry of the catchment area and its morphology: the left bank to the west is broad, gently sloping and heavily cultivated, and differs from the narrow, steeply sloping right bank to the east, which primarily comprises forest and grassland (figure 9). The catchment area can be seen as exclusively agricultural; there are no industries and no settlements of over 200 inhabitants. Agriculture is organized in maize monoculture or mixed polyculture-animal production systems. This has resulted in the area being split into often small, irregular-shaped plots.

The main land cover types seen in the study zone in 1996 were, in decreasing order of area: grassland, forest, maize and cereals, followed by sunflower, fallow, buildings, soybean and rapeseed (Colin 2002).

[Insert figure 9 about here]

4.2. *Presentation of the classified SPOT image*

The image's radiometric information was obtained from the three SPOT channels: XS1 (green channel, from 0.50 to 0.59 µm), XS2 (red channel, from 0.61 to 0.68 µm) and XS3 (near-infrared channel, from 0.79 to 0.89 µm).

From the initial population comprising pixels from the SPOT Image, we chose only pixels whose edges were at least 20 m from the edge of a plot, by superimposition over a vector database corresponding to the agricultural plot structure. We thus assumed that the remaining population (about 85 % of the initial population) contained few mixed pixels and that the effect on radiometry of georeferencing the image was limited.

The cover types under-represented in the catchment area were eliminated due to the lack of representativeness of the samples. These were tobacco, market garden crops, water, soybean, vineyards and buildings. In view of cropping schedules and of the date of the image, the chosen types were grouped together, resulting in a nomenclature comprising five main classes: forest (class A), maize (class B), cereals and rapeseed (class C), grassland-fallow-sorghum (class D) and sunflower (class E).

4.3. *Study of part of the catchment area*

4.3.1 *Description of the zone*

From this area, we extracted an initial window representative of land cover in the downstream section, representing around 10% of the total catchment area (see extract in figure 9). The geographical distribution of the different classes was quite marked in this zone.

[Insert figure 10 about here]

[Insert figure 11 about here]

The zone contained 4944 pixels (figure 10). Each pixel was characterized by its three channels (XS1, XS2 and XS3), its coordinates (Xcoord and Ycoord) and its land cover class.

In this case, exhaustive information was available on all these aspects. Pixel distribution in the area (XS2-XS3) is shown in figure 11. Table 7 gives a breakdown of the different classes and the numbers involved.

[Insert table 7 about here]

The set of data was split into three, by random drawing, without replacement: a basic sample and a test sample, each comprising 247 pixels (i.e. a 5% sampling rate), with the remainder (4450 pixels) making up an anonymous sample.

4.3.2 *Results of boosting*

In this case, the boosting operation took account not only of the three channels, XS1, XS2 and XS3, but also of the geographical coordinates, Xcoord and Ycoord.

Percentage of well classified items

The algorithm achieved 100% well classified items for the basic sample at step 99, while the maximum percentage of well classified items for the test sample was 87.9% at step 56 (figure 12) .

[Insert figure 12 about here]

Boosting weights: these were studied for each class up to step 99. We shall present only the results for the maize class. In this class, pixel 11- was seen to have a high weight (figure 13). It could be represented within both the XS2-XS3 radiometric variable area (figure 14) and the geographical area (figure 15) and was located on the edge of both areas.

[Insert figure 13 about here]

[Insert figure 14 about here]

[Insert figure 15 about here]

Importance of variables: we calculated the contribution of the variables to the establishment of learners during the first 56 steps. This showed that the contribution of the geographical coordinates was around 55% (table 8).

[Insert table 8 about here]

Mapping learner values for each class: we have seen that at the end of each step t and for each class, boosting provided the value of the learner that could be used to class the pixels from the basic, test and anonymous samples. As the value of the learner is interpreted according to its sign and absolute value, we were able, by referring to the geographical area, to identify zones whose likelihood of classification in a given group was high, average or low. Figure 16 below reveals zones with a strong chance of being classified in the forest class (red), others with only a small chance of being classified in the forest class (blue) and others for which there was a degree of uncertainty (yellow).

[Insert figure 16 about here]

Mapping pixels reclassified by boosting: to compare our results with the initial image for which the relevant information was known, we visualized in figure 17 the pixels reclassified by boosting. This enabled us to identify wrongly classified zones, which were often in borderline areas.

[Insert figure 17 about here]

4.4. ***Study throughout the catchment area***

Using a 5% sampling rate, by random drawing without replacement, for the basic and test samples, we obtained a percentage of well classified items for the test sample of 77.7% at step 677. We saw that the geographical structure was much less marked in the catchment area as a whole (figure 9).

4.5. ***Comparison with the maximum likelihood method*:**

For the same samples with the same variables (XS1, XS2, XS3, X, Y) , we produced a classification with the maximum likelihood method, without taking account of a priori information. We obtained the following results (see table 9 below):

[Insert table 9 about here]

Boosting again proved to be a better classifier than maximum likelihood. The difference was less marked when data was taken from the whole catchment area, but there was still a difference of around 7% between the two methods.

The poor results for MLC (maximum likelihood classification) on the part of the catchment area can be explained by the spatial structures of the classes at this spatial scale. This spatial structures present shapes in X,Y space that are captured by boosting. Theses shapes cannot be fitted by the ellipses due to the hypothesis of normality for distributions of the MLC model.

For computational aspects, the application of the boosting algorithm on the whole catchment samples took from 2 to 15 seconds on a conventional Pentium PC.

5. Conclusion

The results obtained for theoretical examples showed that boosting was equivalent to the maximum likelihood classification method when geographical coordinates were not taken into account.

However, for both the theoretical and the studied cases, boosting had the advantage, except if the radiometric variable classes did not overlap.

This advantage was largely due to the fact that boosting may use variables that present complex shapes in the variables hyperspace (as geographical space when there is a minimum of spatial structuring).

The flexibility of the algorithm also makes it possible to take account other information likely to improve classification that may be available for the image as a whole (land use maps, etc.). Another positive aspect of boosting lies in the possibility of mapping criteria for the quality of the classification of each pixel. This map would thus make it possible to qualify the classification by identifying three types of pixels in each group ℓ: those for which there is a

high certainty of classification in group ℓ, those for which there is a **high certainty of not being classified** in group ℓ, and lastly those whose attribution to group ℓ is **uncertain**.

Boosting does not estimate parameters for the classification model compared to MLC. Consequently, it's less sensitive to over-parameterisation while data dimensionality increases with same amount of samples.

While boosting produces better results, it should be used with caution : it is essential to use the test sample to check the decision-making rule obtained (as boosting is highly dependent on the basic sample), and weights need to be checked to detect possible outliers or intruders. One of the main weaknesses of the method is also that it does not enable an explanatory approach: it does not seek to find the variables that explain the difference between groups. However, certain explanatory criteria can be provided, for instance the number of times that each variable intervened in iterations, which should make it possible to determine the importance of the variables.

In the geomatics field, the simplicity of the decision-making rules makes boosting an ideal decision support tool that can easily be integrated into GIS. The final decision-making rule, which is simply a weighted linear combination, can easily be programmed into most GIS.

References

ARNAUD, M., BAILLY, J.S., BOURGEON, G., and PUECH, C., 2002, Le Boosting : une méthode de classification non paramétrique. Revue internationale de géomatique. Volume 12 – n° 4/2002 (under press).

BISHOP, C., 1995, Neural Networks for Pattern Recognition, Clarendon Press, Oxford.

BLAKE, C.L., and MERZ, C.J., 1998, UCI Repository of machine learning databases [http://www.ics.uci.edu/~mlearn/MLRepository.html]. Irvine, CA: University of California, Department of Information and Computer Science.

BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R. A., and STONE, C. J., 1984, CART. Classification and regression trees, Wadsworth and Brooks/Cole (Advanced books and Software, Monterey, California).

BRIEM, G. J., BENEDIKTSSON, J. A., and SVEINSSON, J. R., 2001, Boosting, bagging and consensus based classification of multisensor remote sensing data. 2nd Workshop on Multiple Classifier Systems. Cambridge, July 2001.

COLIN, F., RACLOT, D., and PUECH, C., 2002. Traitement d’image en mode parcellaire couplé à un système de règles de décision Application à la classification de l'occupation du sol en zone de petit parcellaire agricole. Revue Internationale de Géomatique (under press).

DIETTERICH, T. G., 1999a, Machine learning research: Four current directions. AI Magazine, 18(4), pp. 97-136.

DIETTERICH, T. G., 1999b, An experimental comparison of three methods for constructing ensembles of decision trees : Bagging, Boosting and randomization.. Machine learning. pp. 1-22.

DRUCKER, H., CORTES C., 1996, Boosting decision trees. In Advances in Neural Information Processing Systems 8, pp. 479-485.

DUFOUR, A., 2001. Méthode de reconnaissance de l'occupation du sol par image satellite et système expert en petit parcellaire. Application au bassin versant du Sousson (Gers). DEA BEE, Univ. Montpellier II.

FREUND, Y. and SCHAPIRE, R. E. 1995, A decision-theoric generalization of on-line learning and an application to Boosting. Tech. rep., AT&T Bell Laboratories, Murray Hill, NJ.

FREUND, Y. and SCHAPIRE, R. E., 1996, Experiments with a new Boosting algorithm. In Saitta, L. (Ed.), Proceedings of the Thirteenth International Conference on Machine Learning, San Francisco, CA. Morgan Kaufmann. pp. 148-156.

FRIEDMAN, J., HASTIE, T., and TIBSHIRANI, R., 1998, Additive Logistic Regression : a Statistical View of Boosting. Technical Report.

HASTIE, T., TIBSHIRANI, R., and FRIEDMAN, J., 2001, The Elements of Statistical Learning – Data Mining, Inference, and Prediction. Springer-Verlag.

MACLIN, R., and OPITZ, D., 1997, An empirical evaluation of bagging and boosting. In Proceedings of the Fourteenth National Conference on Artificial Intelligence. pp. 546-551.

MONGET, J.M., 1997, Classification Techniques in Remote sensing and cartography, Ecole des Mines de Paris, 100 p.

QUINLAN, J. R., 1993, C4.5 : Programs for Empirical Learning. Morgan Kaufmann. San Francisco, CA.

QUINLAN, J. R., 1996, Bagging, Boosting and C4.5 : In Proceedings of Thirteenth National Conference on Artificial Intelligence. Cambridge, MA. AAAI Press/MIT Press. pp. 725-730.

RIPLEY, B.D., 1996, Pattern Recognition and Neural Networks, Cambridge University Press.

SCHAPIRE, R.E., 1999, Theoretical views of boosting and applications. In Tenth International Conference on Algorithmic Learning Theory.

SCHAPIRE, R. E., and SINGER, Y., 1999, Improved boosting algorithms using confidence-rated predictions, Machine Learning, 37, 297-336.

SCHAPIRE, R. E., and SINGER, Y., 2000, BoosTexter: A boosting-based system for text categorization. Machine Learning, 39, 135-168.

SCHOWENGERDT, R. A., 1996, Techniques for image processing and classification in remote sensing, Ed. Academic press.

XIUPING, J., RICHARDS, J. A., and RICKEN, D. E., 1999, Remote Sensing Digital Image Analysis: An Introduction, Ed. Springer Verlag.

YANG, Y., 1999, An evaluation of statistical approaches to text categorization. Information Retrieval.

1. Assuming :

- $\{(\mathbf{x_1}, y_1),\ldots,(\mathbf{x_m}, y_m)\}$ is an initial set, where:

$\forall$i=1,…,m: $\mathbf{x_i} \in \mathbf{R}^p$ and $y_i \in Y = \{1,\ldots,K\}$;

- $\{(\mathbf{x'_1}, y'_1),\ldots,(\mathbf{x'_n}, y'_n)\}$ is a test set, where:

$\forall$i=1,…,n: $\mathbf{x'_i} \in \mathbf{R}^p$ and $y'_i \in Y = \{1,\ldots,K\}$;

- T is the number of trials;
- $D_t(i,\ell)$ is the weight of observation i for group ℓ at step t;
- For t=1 and $\forall$i=1,…,m: $D_1(i,\ell) = 1/mK$ (initial weights).

2. Repeat for t = 1 to T:

(a) Using weights $D_t(i,\ell)$, find from the initial set the appropriate weak learner h_t:

$$h_t: \mathbf{R}^p x Y \to \mathbf{R} ;$$

(b) Choose coefficient $\alpha_t \in \mathbf{R}$;

(c) Calculate:

$\forall$i=1,…,m and $\forall\ell$=1,…,K : $D_{t+1}(i,\ell) = \dfrac{D_t(i,\ell)\exp\{-\alpha_t\ y_i(\ell)h_t(\mathbf{x_i},\ell)\}}{Z_t}$

where: Z_t is a normalization factor chosen such that D_{t+1} is a distribution;

and $y_i(\ell) = \begin{cases} +1 & \text{if } y_i = \ell \\ -1 & \text{if } y_i \neq \ell \end{cases}$

(d) calculate the % of well classified items for the basic sample and the test sample using the classifier associated with $\sum_{i=1}^{t} \alpha_i h_i$.

3. This produces the final classifier H(**x**) associated with weak learner $\sum_{i=1}^{T} \alpha_i h_i$:

$$H(\mathbf{x}) = \{\ell \in Y = \{1,\ldots,K\} \text{ such that } \sum_{t=1}^{T} \alpha_t h_t(\mathbf{x},\ell) \text{ maximum}\}$$

Table 1: *AdaBoost.MH algorithm (Schapire and Singer 1999)*

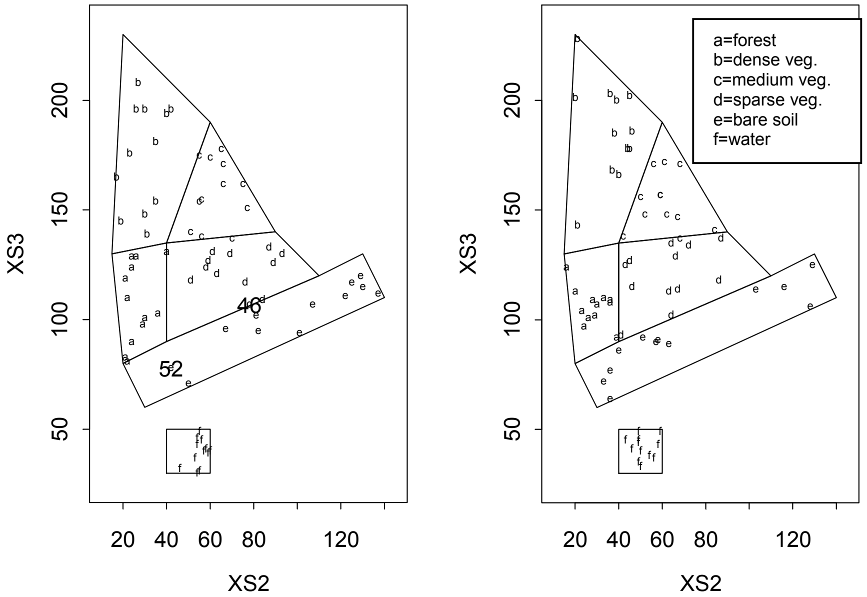

Figure 1: *Pixel position within the radiometric area (N = 12): left, basic sample; right, test sample*

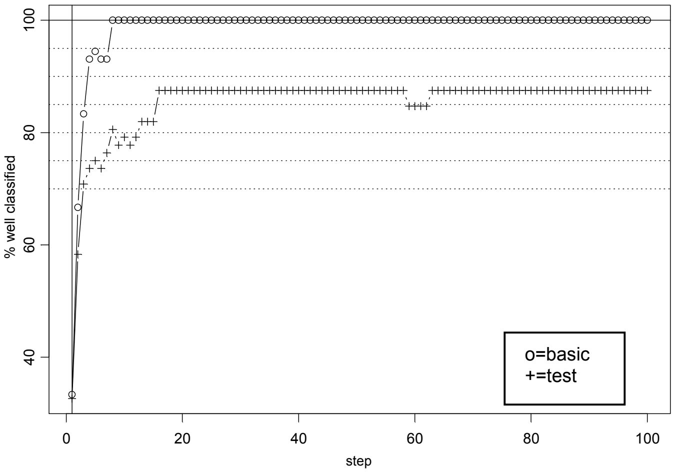

Figure 2: *Changes in the percentage of well classified items for the basic and test samples (N = 12)*

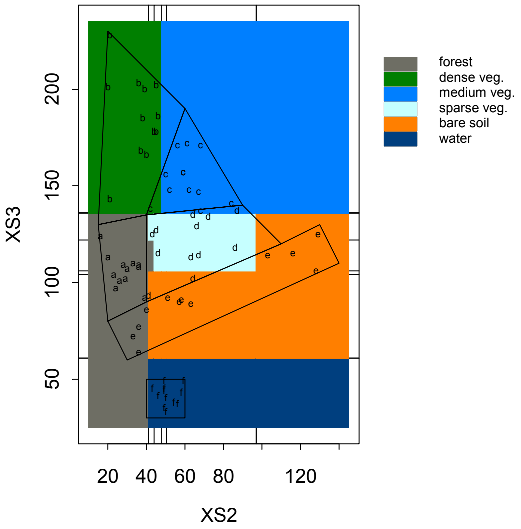

Figure 3: *Decision area after step 16 (N = 12) and test sample pixels*

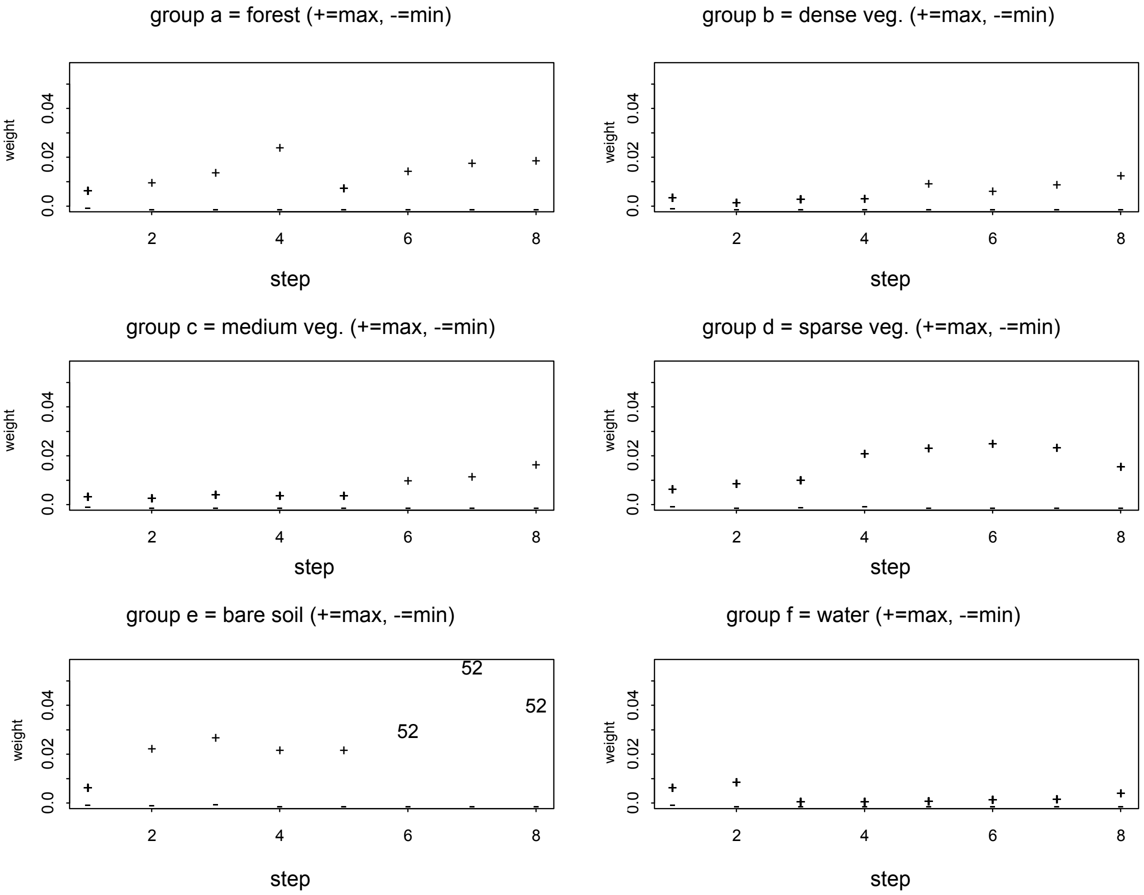

Figure 4: *Study oj tne changes in boosting weights in each group up to step 8 (N = 12)*

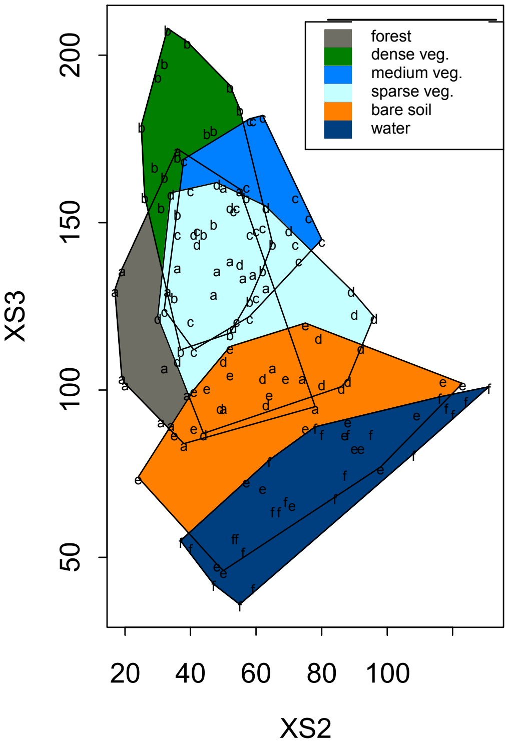

Figure 5: *Convex envelope of radiometric values for each class, for the basic and test samples (N = 12)*

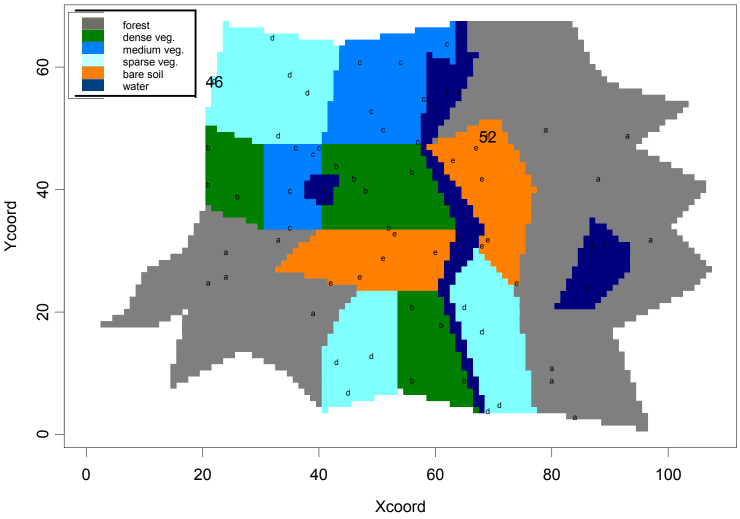

Figure 6: *Position of the pixels in the basic sample within the geographical area (N=12)*

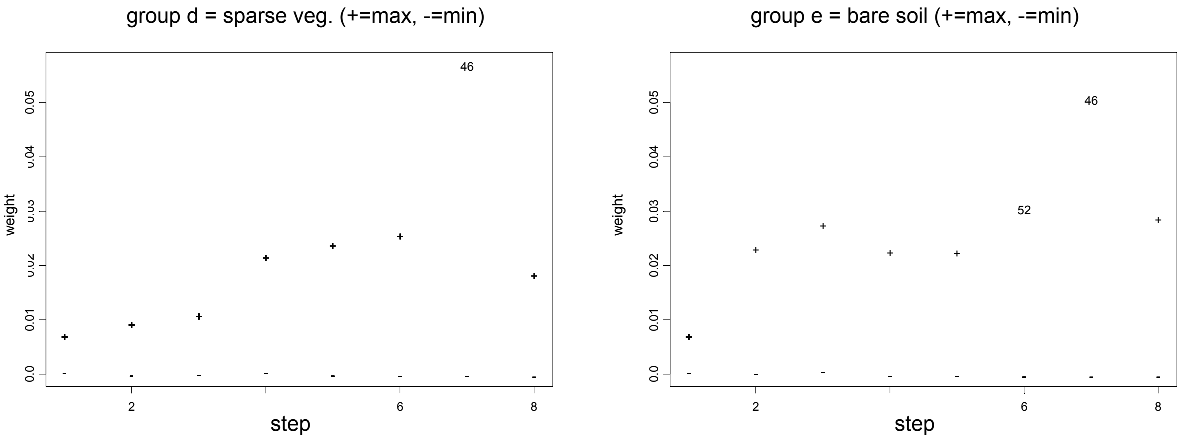

Figure 7: *Changes in boosting weights up to step 8. Left, class 'd'; right, class 'e' (N = 12)*

	N=12 - step 46				N=24 - step 16				N=36 - step 93			
	XS1	XS2	Xcoord	Ycoord	XS1	XS2	Xcoord	Ycoord	XS1	XS2	Xcoord	Ycoord
Number	15	21	4	6	5	8	1	2	23	35	20	15
%	32.6	45.7	8.7	13.0	31.3	50.0	6.2	12.5	24.8	37.6	21.5	16.1

Table 2: *Contribution of variables for N = 12, 24 and 36 in the case of a non-random spatial structure and without radiometric variable overlapping*

	N=12 - step				N=24 - step				N=36 - step			
	XS1	XS2	Xcoord	Ycoord	XS1	XS2	Xcoord	Ycoord	XS1	XS2	Xcoord	Ycoord
Number	13	11	12	13	8	12	16	9	6	8	10	6
%	26.5	22.5	24.5	26.5	17.8	26.7	35.5	20.0	20.0	26.7	33.3	20.0

Table 3: *Contribution of variables for N = 12, 24 and 36 in the case of a non-random spatial structure and with radiometric variable overlapping*

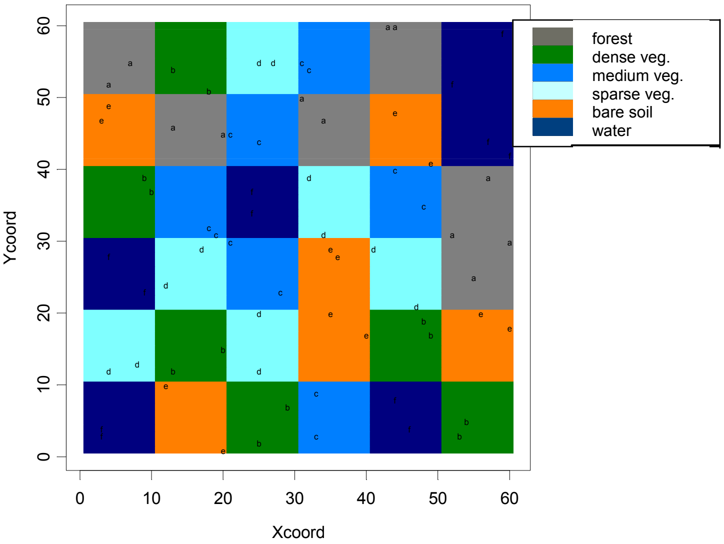

Figure 8: *Representation of the pixels of each class in the case of a random structure (N = 12)*

	N=12 - step 16				N=24 - step 57				N=36 - step 94			
	XS1	XS2	Xcoord	Ycoord	XS1	XS2	Xcoord	Ycoord	XS1	XS2	Xcoord	Ycoord
Number	7	9	0	0	20	29	5	3	27	41	11	15
%	44.0	56.0	0	0	35.1	50.9	8.8	5.2	28.7	43.6	11.7	16

Table 4: *Contribution of variables for N = 12, 24 and 36 in the case of a random spatial structure and without radiometric variable overlapping*

	N=12 – step 40				N=24 - step 66				N=36 - step 98			
	XS1	XS2	Xcoord	Ycoord	XS1	XS2	Xcoord	Ycoord	XS1	XS2	Xcoord	Ycoord
Number	11	11	5	13	12	18	15	21	19	25	28	26
%	27.5	27.5	12.5	32.5	18.2	27.3	22.7	31.8	19.4	25.5	28.6	26.5

Table 5: *Contribution of variables for N = 12, 24 and 36 in the case of a random spatial structure and with radiometric variable overlapping*

		N	Boosting			Without using geographical coordinates
			Without using geographical coordinates	Using geographical coordinates		Maximum likelihood
				Spatial structure	Random structure	
radiometry	without overlapping	12	87.5	87.5	87.5	**95.8**
		24	89.6	**93.8**	89.6	**93.8**
		36	92.1	**97.2**	93.5	95.4
	with overlapping	12	50.0	**69.4**	62.5	41.7
		24	50.7	**81.9**	69.4	48.6
		36	48.6	**82.9**	76.9	47.2

Table 6: *Comparison of the two methods for the test sample*

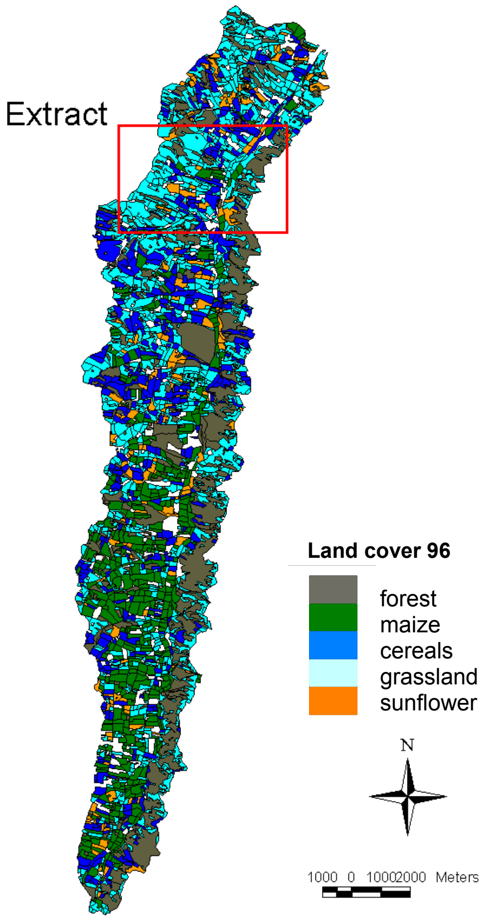

Figure 9: *Land cover in the Sousson catchment area (Summer 1996)*

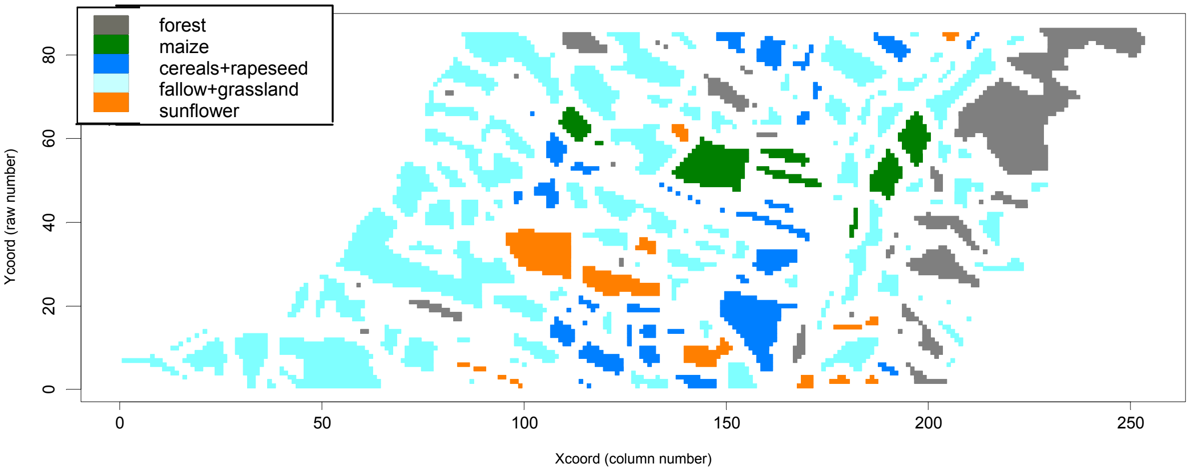

Figure 10: *Visualization of the zone (all pixels)*

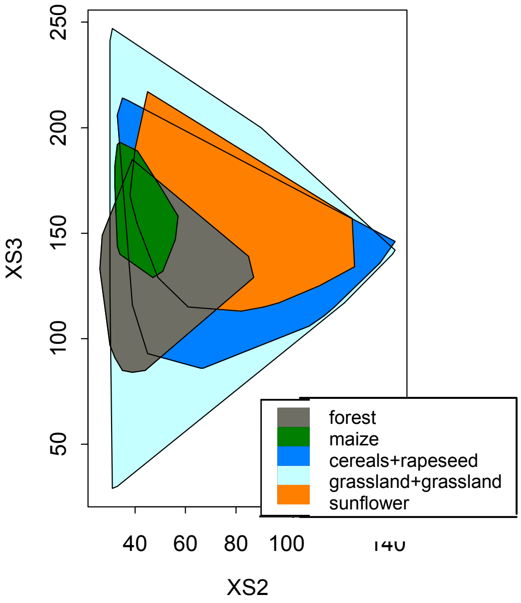

Figure 11: *Convex envelope within the XS2-XS3 radiometric area for the 4944 pixels as a whole*

Type	**Class**	**Numbers**
Forest	A	928
Maize	B	361
Cereals-rapeseed	C	593
Grassland-fallow-sorghum	D	2718
Sunflower	E	344
Total		4944

Table 7: *Pixel distribution according to class*

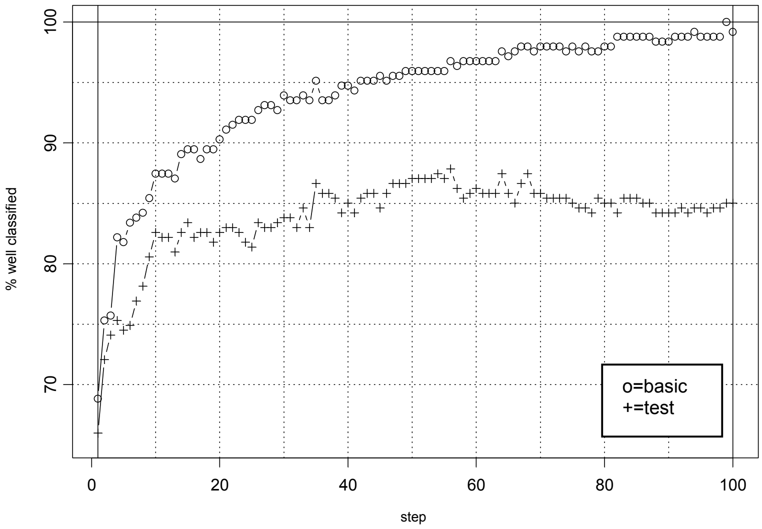

Figure 12: *Percentage of well classified items for the basic and test samples, over the first 100 steps*

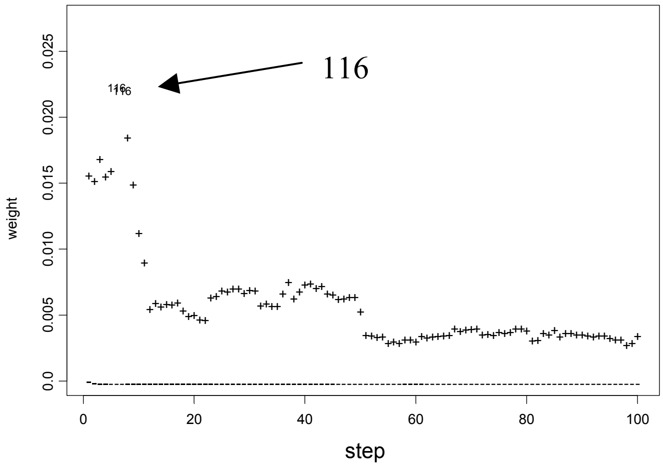

Figure 13: *Identification oj pixel 116, which had a high weight in the maize class*

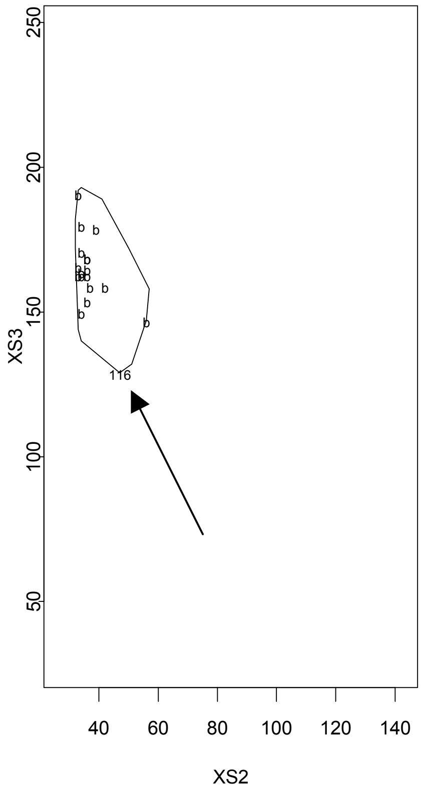

Figure 14: *Identification within the radiometric area of pixel 116, which had a high weight in the maize class*

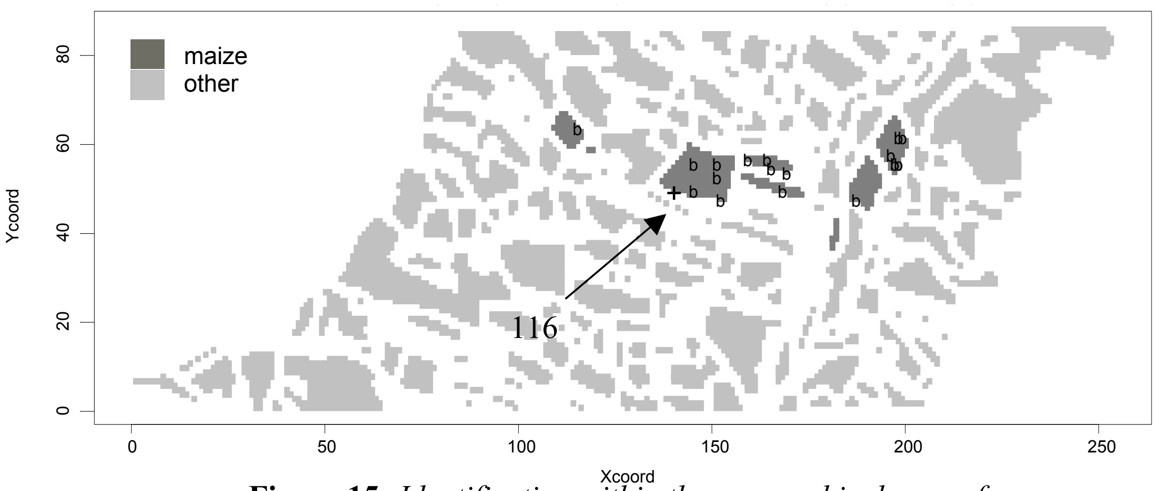

Figure 15: *Identification within the geographical area of pixel 116 (+), which had a high weight in the maize class*

	XS1	XS2	XS3	Xcoord	Ycoord
Number	7	5	3	18	13
%	12.5	8.9	23.2	32.2	23.2

Table 8: *Contribution of variables*

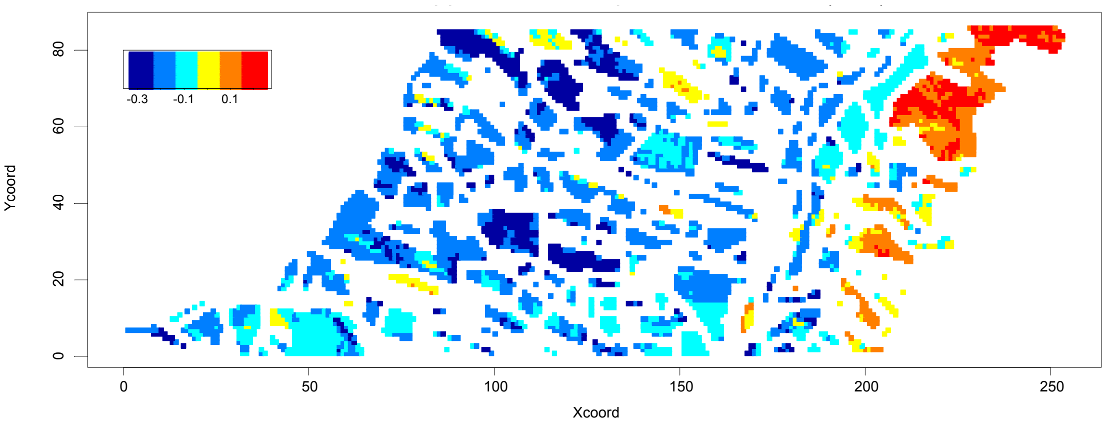

Figure 16: *Map of the values of the weak learner for the forest class within the geographical area*

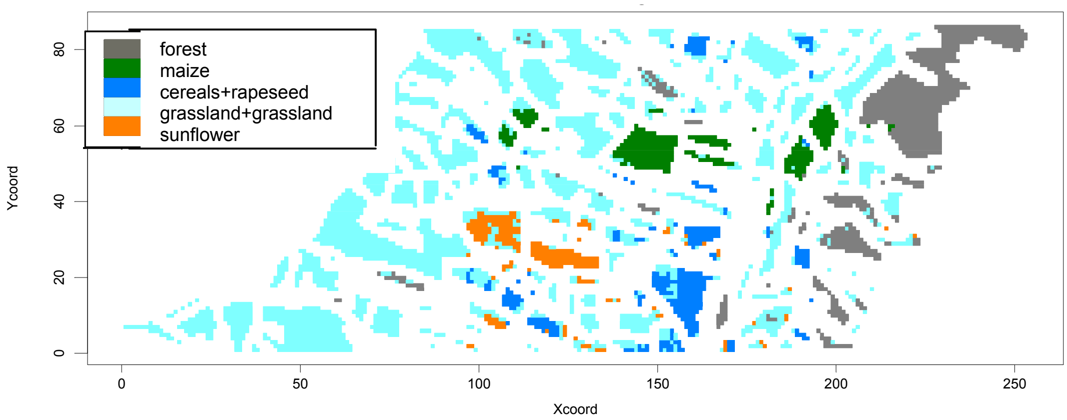

Figure 17: *Map of the new classification produced by boosting at step t=56*

	Boosting	Maximum likelihood
Part of catchment area	**87.9%**	57.9%
Whole catchment area	**77.7%**	70.0%

Table 9: *Comparison of boosting and maximum likelihood*