Accuracy of Statistical Classification Strategies in Remote Sensing Imagery

Alejandro C. Frery
Instituto de Computação
Universidade Federal de Alagoas
57072-970 Maceió, AL – Brazil
acfrery@pesquisador.cnpq.br

Susana Ferrero
Departamento de Matemática
UNRC Ruta 36 km 601
X5804BYA Argentina
sferrero@exa.unrc.edu.ar

Oscar H. Bustos
FaMAF–UNC
Av. Medina Allende s/n
5000 Córdoba, Argentina
bustos@mate.uncor.edu

Abstract

We present the assessment of two classification procedures using a Monte Carlo experience and Landsat data. Classification performance is hard to assess with generality due to the huge number of variables involved. In this case we consider the problem of classifying multispectral optical imagery with pointwise Gaussian Maximum Likelihood and contextual ICM (Iterated Conditional Modes), with and without errors in the training stage. Using simulation the ground truth is known and, therefore, precise comparisons are possible. The contextual approach proved being superior than the pointwise one, at the expense of requiring more computational resources, with both real and simulated data. Quantitative and qualitative results are discussed.

1. Introduction

The production of thematic cartography is one of the main goals of remote sensing image processing and analysis. The aim of this task is producing a map of (possibly all) homogeneous classes present in the scene. This can be achieved by means of visual inspection and manual labour, but digital techniques are more used every day since remote sensing is widely recognized as the primary source of information for cartography.

A thematic map or the overall inventory of classes can be obtained with classification techniques. Such products are essential in many applications as, for instance, crops statistics, mining and hydrological resources studies.

Among the classification strategies [8, 12, 14], we deal with two notorious members of the class of statistical supervised techniques. These procedures have three basic steps, namely training, production and testing, and certain rules have to be obeyed in order to generate good products.

The multivariate Gaussian law is the most widely used distribution for modelling image data acquired with optical remote sensing instruments [16]. This will be the statistical framework employed in this work, since techniques based on this distribution are available in most remote sensing image processing platforms. Though the data exhibit a noticeable spatial structure, that turns into what we perceive as texture, the spatial correlation will not be considered here.

The spatial correlation of the classes is an important source of information, specially when high resolution data is used. This structural information will be modelled using a Markov Random Field description for the (unobserved) classes under estimation. Statistical classification using this model does not become a simple geometric decision rule in the features space; it becomes, in non-trivial cases, a NP-complete problem [9].

Among the approximation algorithms for obtaining a solution to this problem, we chose to work with the ICM (*Iterated Conditional Modes*), which is among the fastest alternatives being the others computing the MAP (*Maximum a posteriori*, that requires simulated annealing) and MPM (*Maximum Posterior Marginals*, obtainable only by costly stochastic simulation) estimators [4].

Using spatial information yields, in principle, better results than using only spectral (pointwise) evidence, but the computational requirements increase. The purpose of this paper is assessing the precision of products obtained by Maximum Likelihood (ML) and by ICM classifications under different training situations, namely, with and without errors. In both cases the data were described by the multivariate Gaussian distribution.

Training is subjected to errors, since it depends on multiple sources of possibly vague and contradictory information (visual analysis, previous experience, data acquired by other sensors and in a different moment etc.). Modelling this error and assessing its influence on supervised classification algorithms is another contribution of this paper.

A Monte Carlo experience was devised in order to carry out this study. Images are simulated and they are automatically classified. In doing so, one has the ground truth before which the results can be compared.

After making this simulation-based assessment, the techniques are applied to real data: more than fifty samples from a Landsat ETM+ image with many thematic classes. All available bands were employed in all but one situation, being this last one designed to evaluate the impact of partial information on the classification procedures.

The rest of the paper unfolds as follows. Section 2 recalls the basic definitions of statistical classification. Section 3 presents the Monte Carlo experiences and the simulation results. Section 4 shows the results of applying the techniques to a set of real data. Finally, section 5 comments the results and their consequences.

2. Supervised Statistical Classification

From a mathematical standpoint, a multispectral image is a three-dimensional real matrix:

$$\mathbf{z} = [z(i,j,k)]_{0\leq i\leq M-1, 0\leq j\leq N-1, 0\leq k\leq K-1}\,, \\ z(i,j,k)\in\mathbb{R}.$$

The two first dimensions are related to the geographical coordinates of the scene, and determine the size of the support of the image ($M \times N$), while the latter is related to the spectral dimension of the data. We say that the image has M columns, N rows and K bands, and $\boldsymbol{z}(i,j)$ will denote, for short, the K dimensional vector observed in site (i,j). The coordinates can also be dropped for the sake of compactness.

A classification rule is a function that, using the availabe information, defines a set of $M \times N$ labels, say $\boldsymbol{c} = [c\,(i,j)]_{0\leq i\leq M-1, 0\leq j\leq N-1}$. This is not a real matrix, but it is defined on a set of L possible labels $\boldsymbol{C} = \{c_1, \dots, c_L\}$. This object should not be regarded as an image, but as a thematic map.

Supervised statistical classification procedures consist of providing such a rule by means of decisions based on the statistical properties of the data, i.e., parameters to be estimated. The steps that these procedures are based upon are those commented before: training, production and testing.

2.1. Multivariate Gaussian Model

This model assumes that the observations related to each of the L classes obey different probability laws characterized by the probability density function

$$f_\ell(\boldsymbol{z}) = \frac{\exp\Big(-\frac{1}{2}\left((\boldsymbol{z}-\boldsymbol{\mu}_\ell)^t \boldsymbol{M}_\ell^{-1}(\boldsymbol{z}-\boldsymbol{\mu}_\ell)\right)\Big)}{(2\pi)^{K/2}(\det(\boldsymbol{M}_\ell))^{1/2}},$$

where K is the number of bands, $\boldsymbol{\mu}_\ell$ is the vector of means and $\boldsymbol{M}_\ell$ is the covariance matrix and $1 \leq \ell \leq L$ is the class index. This assumption is usually verified in practice, if a careful choice of classes is made.

The classification rule that stems from this assumption and the hypothesis of independence among different sites is assigning the site (i,j) to class c_ℓ if

$$f_\ell(\boldsymbol{z}(i,j)) \geq f_{\ell^*}(\boldsymbol{z}(i,j)), \tag{1}$$

for every $1 \leq \ell^* \leq L$, i.e., if the likelihood of the observation $\boldsymbol{z}(i,j)$ is maximized by the model of class ℓ. As stated, this procedure assumes that all classes have the same *a priori* probability. In a Bayesian context, this is equivalent to assigning a non-informative prior to the classes.

In [8] it is shown that this rule is equivalent to decision regions in the features domain in the form of hyperquadrics; other distributions induce different hypersurfaces. It is also easy to see that the rule formulated in equation (1) is equivalent to assigning the site (i,j) to class c_ℓ if

$$-\ln(\det(\boldsymbol{M}_\ell)) - (\boldsymbol{z}(i,j)-\boldsymbol{\mu}_\ell)^t\boldsymbol{M}_\ell^{-1}(\boldsymbol{z}(i,j)-\boldsymbol{\mu}_\ell) \geq \\ -\ln(\det(\boldsymbol{M}_{\ell^*})) - (\boldsymbol{z}(i,j)-\boldsymbol{\mu}_{\ell^*})^t\boldsymbol{M}_{\ell^*}^{-1}(\boldsymbol{z}(i,j)-\boldsymbol{\mu}_{\ell^*}),$$

for every $1 \leq \ell^* \leq L$.

In most practical situations one has to estimate the parameters $\boldsymbol{\mu}_\ell$ and $\boldsymbol{M}_\ell$ using training samples.

2.2. Markov Random Fields

Real data exhibit a great deal of spatial correlation, and this issue is notorious with high resolution imagery. This is due to two main reasons: firstly, the spectral information results from the integration of many sources, including neighboring sites; secondly, classes tend to appear in spacial clusters, e.g. if the true class of a site is “water” it is likely that its neighboring sites are of the same type.

Spatial information, also known as “context”, can be modelled as correlation structures in the observed data, or as spatial dependence among classes, or both. We chose to work with the second, within a Bayesian framework, through a model that puts more weight on classes that exhibit spatial correlation.

Markov Random Fields, the spacial generalization of Markov chains, have deserved a great deal of attention in the computer vision literature since they were successfully used in image restoration [6, 11]. The interested reader is also referred to [17]; they will be reviewed in the following.

Let $C = \{c_1, \dots, c_L\}$ be the set of classes that describes the ground truth. Each element of the support $S = \{(i,j)\colon 0 \leq i \leq M-1, 0 \leq j \leq N-1\}$ is assigned one of these classes, C^S is the set of functions on S with value in C.

A family $\boldsymbol{V} = \{V_{(i,j)}\colon (i,j) \in S\}$ of subsets of S is a *neighborhood* of S if

1. $(i,j) \notin V_{(i,j)}$ for every (i,j),
2. $(i,j) \in V_{(i',j')} \Leftrightarrow (i',j') \in V_{(i,j)}$, and

3. $S = \cup_{(i,j)\in S} V_{(i,j)}$.

The pair $\mathcal{G} = (S, \boldsymbol{V})$ is a graph.

In this work we will consider the so-called "eight-neighbors structure": $V^2_{(i,j)} = \{(i',j') \in S\colon (i'-i)^2 + (j'-j)^2 \leq 2, (i',j') \neq (i,j)\}$, and $\mathbb{V}^2 = \{V^2_{(i,j)}\colon (i,j) \in S\}$ for every $(i,j) \in S$.

A Random Field with space state C^S is a random matrix $\mathcal{C} = [C(i,j)]_{0\leq i\leq M-1, 0\leq j\leq N-1}$ such that $C(i,j)$ is a random variable with values in C. An outcome of $\mathcal{C}$ is an element of C^S. The probability distribution of $\mathcal{C}$ on C^S will be denoted by $\Pr_{\mathcal{C}}$, i.e., for each $\boldsymbol{c} \in C^S$ one has that $\Pr_{\mathcal{C}}(\boldsymbol{c})$ is the probability that $C(i,j) = \boldsymbol{c}(i,j)$ for every $(i,j) \in S$.

Given a graph $\mathcal{G} = (S, \boldsymbol{V})$ and a Random Field $\mathcal{C}$ with space state C^S, we say that $\mathcal{C}$ is a $\mathcal{G}$-Markov Random Field if for every $\boldsymbol{c} \in C^S$ and every $(i,j) \in S$ holds that

$$\Pr_{\mathcal{C}}(A \mid B) = \Pr_{\mathcal{C}}(A \mid V),$$

where $A = \{\boldsymbol{c}'\colon \boldsymbol{c}'(i,j) = \boldsymbol{c}(i,j)\}$, $B = \{\boldsymbol{c}'\colon \boldsymbol{c}'(i',j') = \boldsymbol{c}(i',j'), \forall (i',j') \neq (i,j)\}$, and $V = \{\boldsymbol{c}'\colon \boldsymbol{c}'(i',j') = \boldsymbol{c}(i',j'), \forall (i',j') \in V_{(i,j)}\}$. In words, the conditional distribution of the random variable at site (i,j) given the observation of all other sites depends only on the observed outcomes at the neighboring sites $V_{(i,j)}$.

A particular Markov Random Field model, namely the Potts model, has been widely used to describe the spatial distribution of classes in thematic maps [5, 6, 17]. In order to define it, consider $\beta \in \mathbb{R}$ a real number. A Random Field $\mathcal{C}$ with space state C^S is a Potts model with parameter β with respect to the graph $\mathcal{G} = (S, \boldsymbol{V}^2)$ if

$$\Pr_{\mathcal{C}}(\boldsymbol{c}) = \frac{Z_\beta(\boldsymbol{c})}{Z_\beta}, \tag{2}$$

where

$$Z_\beta(\boldsymbol{c}) = \exp\left(\beta \sum_{(i,j)\in S} N_{(i,j)}(\boldsymbol{c})\right),$$

with

$$N_{(i,j)}(\boldsymbol{c}) = \#\{(i',j') \in V^2_{(i,j)}\colon \boldsymbol{c}(i',j') = \boldsymbol{c}(i,j)\},$$

and the so-called partition function $Z_\beta = \sum_{\boldsymbol{c}\in C^S} Z_\beta(\boldsymbol{c})$. This model states that the log-probability of observing class c_ℓ at coordinate (i,j) is proportional to β times the number of neighboring sites where class c_ℓ occurred. Positive values of β assign more probability to maps with clusters of same classes. Using this model as a prior distribution leads to a classification rule that takes context into account.

2.3. ICM Algorithm with β unknown

The ICM algorithm is an iterative approach to finding better solutions than those provided by a pointwise procedure, such as pixelwise Maximum Likelihood. It starts with an arbitrary solution and improves it replacing the class in every coordinate by the one that maximizes an objective function that, in turn, comprises two terms: the evidence provided by the data (the information on which Gaussian Maximum Likelihood is based upon) and the evidence provided by the context.

In our implementation, ICM starts with the pointwise Gaussian Maximum Likelihood classification. Then, a new classification is obtained using, for every $(i,j) \in S$, $1 \leqslant \ell \leqslant L$, $\boldsymbol{z} \in \mathbb{R}^K$, the following decision rule:

$$\begin{aligned} g_\ell((i,j), \boldsymbol{z}, \boldsymbol{c}, \beta) = \\ \frac{1}{2}\big(-\log(\det(\boldsymbol{M}_\ell)) - (\boldsymbol{z} - \boldsymbol{\mu}_\ell)^t \boldsymbol{M}_\ell^{-1} (\boldsymbol{z} - \boldsymbol{\mu}_\ell)\big) + \\ \beta\#\{(i',j') \in V^*_{(i,j)}\colon \boldsymbol{c}(i',j') = c_\ell\}, \end{aligned} \tag{3}$$

where $V^*_{(i,j)} = V_{(i,j)} \cup \{(i,j)\}$. The first term of the second member in equation (3) is the same as the pointwise Maximum Likelihood classification rule under the Multivariate Gaussian model. The second term is the contextual component that, provided $\beta > 0$, puts more weight on those classes that surround site (i,j).

The contextual influence is quantified by the value of the parameter β. When $\beta = 0$ the rule provided by equation (3) reduces to pointwise Maximum Likelihood classification under the Multivariate Gaussian model, i.e., context has no effect on the evidence provided by the observed data; when $\beta \to \infty$ the effect is reversed, i.e., the observed data have no influence on the rule, which is solely the local mode. This parameter is unknown and, therefore, is has to be informed.

The literature reports implementations where the value of β is provided by means of trial-and-error procedures [2]. Our approach consists of estimating it from the available information by pseudolikelihood [3]: since maximum likelihood is not feasible due to the cumbersome form of equation (2), it is replaced by the product of conditional laws. This estimation is performed after each iteration, being the first classification the Gaussian Maximum Likelihood rule or, equivalently, the rule provided by equation (3) setting $\beta = 0$. An iteration consists of (i) estimating β from the previous classification and (ii) applying the rule provided in equation (3). It is observed that the sequence of estimated parameters is non decreasing, i.e, that $\widehat{\beta}(0) \leq \widehat{\beta}(1) \leq \cdots$, so one will always end up with a classification with more homogeneous patches than the first provided as starting solution.

The algorithm proceeds until evidence of convergence is achieved. In our implementation at least one of two criteria has to be satisfied in order to stop the procedure: a certain maximum number of iterations (fixed in 100 in our experiments) or a certain minimum percentage of classes changed (set to 5%).

2.4. Classification Stages

After image registration, callibration and feature extraction, supervised classification consists of three stages:

Training: the number of distinct classes is identified, and representative samples are collected. Part of these samples are used to estimate the parameters of the spectral signature of each class (training samples), and the rest (test samples) is used to assess the overall accuracy of the procedure. If the multivariate Gaussian distribution is assumed for each class, the vector of means and the covariance matrix are estimated. Wrong choices in this stage will propagate errors in an unpredictable way hampering, thus, the quality of the final product.

Production: Each coordinate is assigned to the class that satisfies a certain decision rule producing, thus, a thematic map.

Testing: the accuracy is estimated checking the class each test sample is associated to. An error matrix (also called confusion matrix) is then built, and associated statistics can be computed such as the Kappa coefficient of agreement.

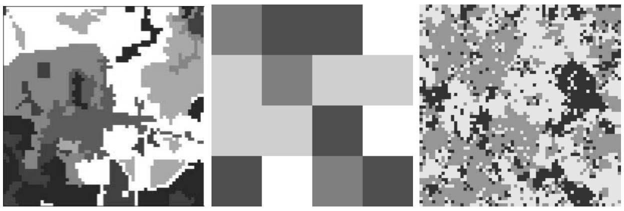
3. Precision Assessment by Simulation

Three types of class images were used in this work in order to describe typical situations that appear in practice: a hand-painted one (called "Cubism", see Figure 1(a)) inspired in thematic maps, random blocks, and outcomes of the Potts model.

Maps in the shape of random blocks with L classes are obtained dividing the support S, which is a square of side 64 or 72 in squares of side 4 or 6, respectively, and drawing a class independently from the other for every small square; if the same class is drawn in every small square, the map is discarded and the procedure begins again. A typical outcome for a 64×64 support and $L = 4$ is shown in Figure 1(b).

Figure 1(c) shows a typical outcome of the Potts models, as defined in equation (2) with four classes and $\beta = 1/2$.

In order to make the assessment in as many as possible representative situations, fourteen situations were considered: the three types of class images of sizes $64 \times 64 \times K$ and $72 \times 72 \times K$, where $K = 3$ or 4 bands and 4 or 6 classes. Besides these models, two training situations were modelled: with and without errors in the training stage. This last parameter in the simulation is of paramount importance since, as will be seen, the quality of the training samples is critical and this issue has not been fully addressed in the literature.



(a) "Cubism" (b) Random blocks (c) Potts model

Figure 1. Images used in the assessment

The parameter values for each situation, i.e., $(\boldsymbol{\mu}_\ell, \boldsymbol{M}_\ell)$, were chosen with the following rules (see details in Appendix A):

P1: Real values (as presented in [18]; see Appendix A).

P2: Three classes with comparable low mean values and three with comparable high mean values; variances for the classes having comparable values are the same.

P3: The same mean for all the classes; covariances are the same, and variances are increasing.

P4: All classes with the same mean values, with increasing variances and covariances.

Table 1 presents the fourteen situations.

Two hundred replications were made in every situation in order to assess the performance of the pointwise and contextual procedures. Each replication consists of assuming a certain image class, sampling from its distribution if it is of type random blocks or Potts model, transforming classes into observations following the assumed models, obtaining samples for each class (with or without errors), producing the two classifications and validating them.

The training stage consists of choosing, for each class, a random sample of sites of size 10% of the observed number of sites of that class, and using the corresponding observations for parameter estimation. In the presence of training errors (situations 4, 6, 8, 9, 10, 12 and 14), $1/10$ of those observations is replaced by data from another class uniformly chosen among the others.

Since the true class image is known beforehand, it is possible to compute the actual error matrix and the coefficients of overall accuracy and Kappa with their respective confidence intervals [1, 7, 10].

The two hundred replications for each situation allow us to draw the following conclusions:

- Most situations produce values of Kappa higher than 0.70, so most classifications can be considered "good".
- The lowest coefficients (overall accuracy and Kappa) were achieved in situations 13 and 14, where there was

Table 1. Parameters for the observations

Situation	Image Type	Size	Classes	$(\boldsymbol{\mu}_\ell, \boldsymbol{M}_\ell)$	Error
1	Blocks	$64 \times 64 \times 4$	4	P1	N
2	Blocks	$72 \times 72 \times 3$	6	P2	N
3	Blocks	$64 \times 64 \times 4$	4	P3	N
4	Blocks	$64 \times 64 \times 4$	4	P3	Y
5	Potts	$64 \times 64 \times 4$	4	P1	N
6	Potts	$64 \times 64 \times 4$	4	P1	Y
7	Potts	$72 \times 72 \times 3$	6	P2	N
8	Potts	$72 \times 72 \times 3$	6	P2	Y
9	Potts	$64 \times 64 \times 4$	4	P3	Y
10	Potts	$64 \times 64 \times 4$	4	P3	Y
11	Cubism	$64 \times 64 \times 3$	6	P2	N
12	Cubism	$64 \times 64 \times 3$	6	P2	Y
13	Cubism	$64 \times 64 \times 3$	6	P4	N
14	Cubism	$64 \times 64 \times 3$	6	P4	Y

a high level of confusion: same mean values for every class and increasing variances and covariances.

- Situations 3 and 4 also produced low coefficients, but in this case ICM doubled the quality of pixelwise classification.
- Coefficients computed on ICM classifications are higher than the others in those situations where training was subjected to error.
- All coefficients are significatively different, and in most cases the evidence provided is that ICM is better than pixelwise classification.

Figure 2 summarizes some of these results, showing the 95% confidence intervals of the Kappa coefficient in some of the simulated situations. Light lines correspond to the Maximum Likelihood algorithm, while thick ones show the results obtained with ICM. It is clear that ICM significantly and consistently outperforms ML.

4. Analysis of a Landsat ETM+ image

Jackson and Landgrebe [13] use an ICM algorithm with fixed values of β, and they show that a contextual classification with small samples attains an accuracy comparable with that obtained with pixelwise maximum likelihood. Arbia et al. [2] also use fixed values of β in a two-class classification setup using simulated data.

In this paper, two experimental setups were considered: one where a 400×233 pixels image was analyzed and other where the whole dataset (6920×5960 pixels) was treated. The first experience aims at assessing the influence of not using all the available information; ML and ICM classifications obtained with the three bands that provide the least separation are compared. In the second setup, 50 subimages were generated from the complete data set in order to make a quantitative comparison of ML and ICM in real situations. We estimate β from the available data. No similar results were found in the literature.

4.1. Setup 1

An area of 400×233 pixels from the 229083 Landsat 7 ETM+ image (30 m resolution) acquired in 2000 over the city of Río Cuarto, Argentina, was analyzed. The 453 RGB composition of the image is shown in Figure 3(a).

Six thematic classes were identified using prior knowledge, exploratory data analysis and photointerpretation: River (predominantly Black in the RGB composition, type # 1, Red in the classification), Urban (Light Blue, # 2, Green), Bare Soil (Light Green, # 3, Blue), Natural Pasture (Dark Green, # 4, Yellow), Managed Pasture (Orange, # 6, Cyan) and Trees (Red, # 6, Magenta).

In order to estimate the vectors of means and the covariance matrices, 5672 training samples were chosen (about 6% of all the pixels). These observations were subjected to a careful exploratory analysis, since the quality of these samples is paramount for obtaining good results. Test samples were also identified in order to assess classification accuracy; in this study 4041 pixels were labeled as test samples.

The reference classification was obtained with the seven available bands by ML; it is shown in Figure 3(b) and its estimated accuracy is 0.86 (see Table 2 for details).

The bands that provided the weakest separation between classes are 1, 3 and 5; the parameters were estimated using this information, and ML classification was obtained (see Figure 3(c). It was then used as the starting point of the ICM algorithm, that ended with $\widehat{\beta}(2) = 0.81$ and the clas-

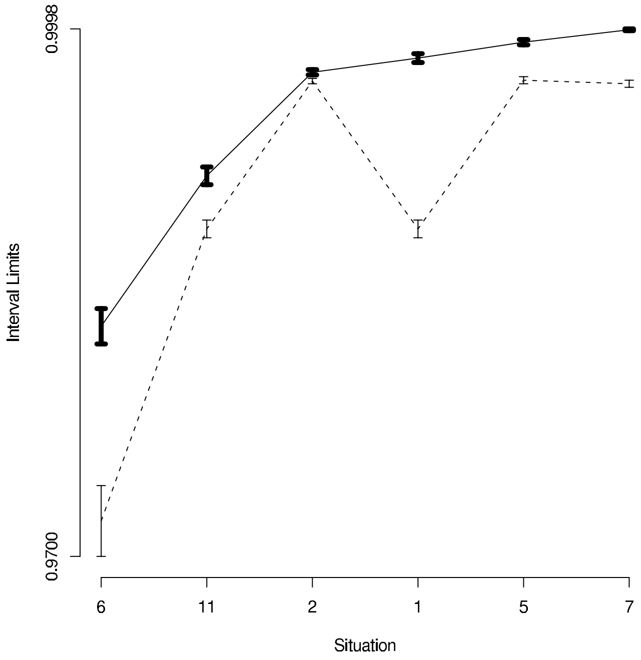


Figure 2. Confidence intervals for Kappa 95%

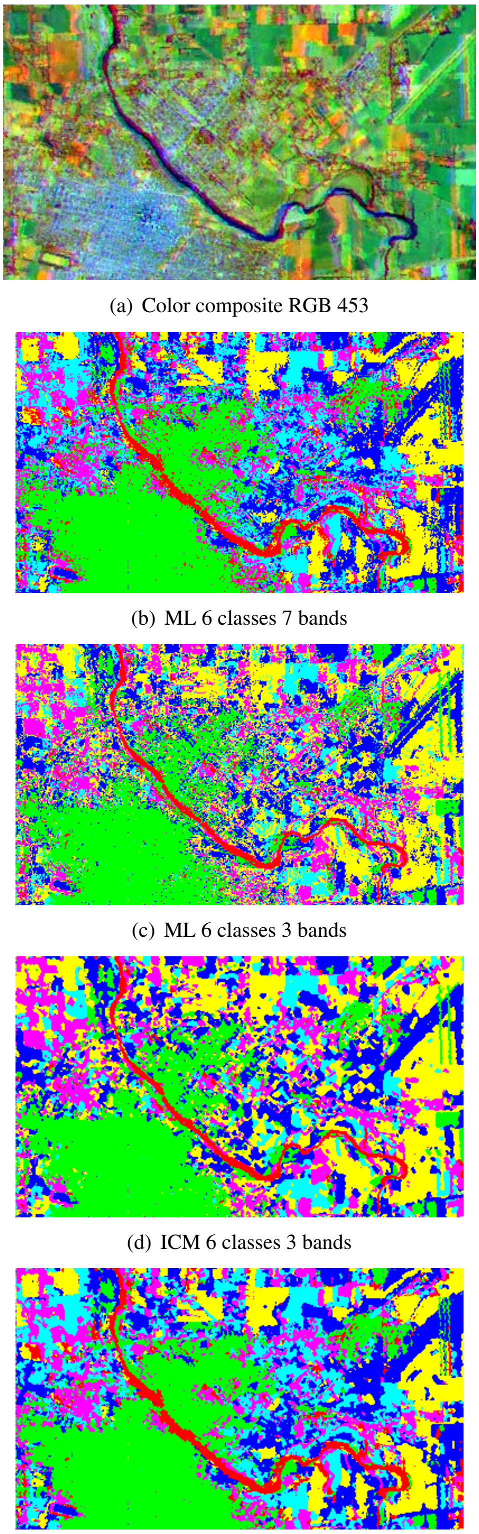
Table 2. Influence of partial information

Technique, data	Acc.	Kappa	95% Conf. Int.
ML, 7 bands	0.86	0.8188	$[0.8049, 0.8328]$
ML, 3 bands	0.79	0.7141	$[0.6976, 0.7306]$
ICM, 3 bands	0.84	0.7872	$[0.7722, 0.8021]$
ICM, 7 bands	0.88	0.8447	$[0.8317, 0.8579]$

sification is shown in Figure 3(d). Quantitative results are shown in Table 2; all estimated accuracy values are high, showing that the classification procedure is excellent [15]. ML with all the available information provides an accuracy of 0.86 but using the three worst bands this figures goes to 0.79; using contextual information on the three worst bands improves the result to 0.84, which is closer to the accuracy obtained with seven bands. Confidence intervals for the accuracy show that these values are significantly different. Incidentally, the accuracy achieved by ICM and seven bands is of 0.88; this classification is shown in Figure 3(e). Figure 3 shows that classifications obtained by ICM are less grainer than those that employed only spectral information.

4.2. Setup 2

From the complete data set (seven bands 6920×5960 image), 50 non-overlapping sub images of size 200×160 with seven bands each were generated. Each was subjected



(a) Color composite RGB 453

(b) ML 6 classes 7 bands

(c) ML 6 classes 3 bands

(d) ICM 6 classes 3 bands

(e) ICM 6 classes 7 bands

Figure 3. Color compositon and maps

to visual and descriptive analysis for the identification of classes, and the number of land covers was 4, 5, 6 or 7. Training and test areas were then selected, with at least 100 sites for each class, and each image was classified by ML and ICM. ICM required at most two iterations, and ended with $\widehat{\beta} \in [0.60; 0.85]$ in all situations, and within the upper half of the interval in 23 out of 50 situations. After classification, the Kappa coefficient of agreement (along with its 90% confidence interval) and accuracy were estimated.

In most situations the coefficient of agreement is close to 1 regardless the classification procedure, so we can conclude that both techniques are in good agreement with the ground truth.

Regarding Kappa, ICM produce equal or better classifications than ML, and in eight out of fifty situations the improvement is statistically significative at the 90% level. Figure 4 shows the estimated values of Kappa for both classification techniques (ML squares and dashed lines, ICM circles and solid lines) in a few situations.

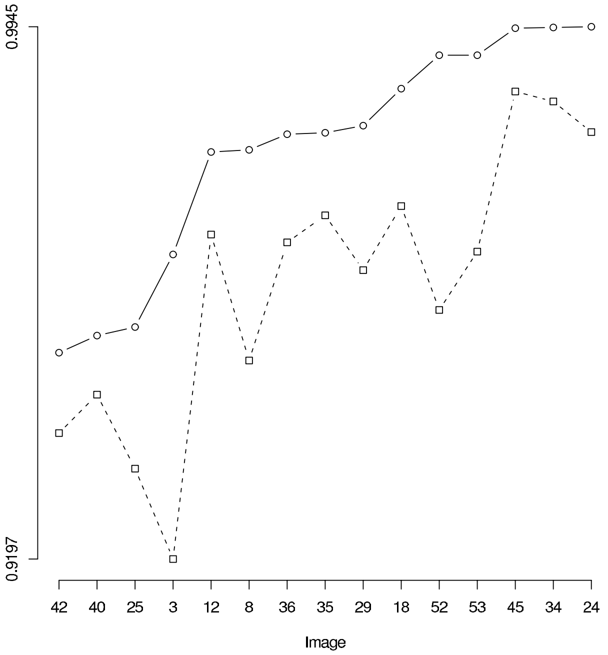


Figure 4. Kappa from ICM and ML maps

5. Results and Conclusions

The precision of two classification techniques in scenarios that include the modelling of user errors in the training stage and a variety of spectral situations was assessed with a Monte Carlo experiment.

Using real data we conclude that (i) training and test samples were carefully chosen, leading to good classification results, and (ii) ICM is always better than ML, but performs the best when there is less than optimal available information compensating the lack of dependable spectral information with contextual evidence.

The evidence collected allows us to say that the ICM contextual classification technique is the most adequate in every situation, even if the training data are collected in a non-dependable manner, so if the computational effort required is not an issue it is always recommended.

References

[1] A. Agresti. *Categorical Data Analysis*. John Wiley & Sons, Inc., 1990.

[2] G. Arbia, R. Benedetti, and G. Espa. Contextual classification in image analysis: An assessment of accuracy of ICM. *Computational Statistcs & Data Analysis*, 30:443 – 455, 1999.

[3] B. C. Arnold and D. Strauss. Pseudolikelihood estimation: some examples. *Sankhya: The Indian Journal of Statistics Series B*, 53:233–243, 1991.

[4] J. Besag. Towards Bayesian image analysis. *Journal of Applied Statistics*, 16(3):395–407, 1989.

[5] O. H. Bustos, A. C. Frery, and S. Ojeda. Strong markov processes in image modelling. *Brazilian Journal of Probability and Statistics- REBRAPE*, 12(2):149–194, 1998.

[6] P. Carnevalli, L. Coletti, and S. Patarnello. Image processing by simulated annealing. *IBM Journal of Research and Development*, 29(6):569–579, Nov. 1985.

[7] R. G. Congalton. A review of assessing the accuaracy of classifications of remotely sensed data. *Remote Sensing of Environment*, 37:35–46, 1991.

[8] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern classification*. John Wiley & Sons, New York, 2 edition, 2001.

[9] P. A. Ferrari, A. Frigessi, and P. G. Sa. Fast approximate MAP restoration of multicolor images. *Journal of the Royal Statistical Society B*, 57(3):485–500, 1995.

[10] R. W. Fitzgerald, , and B. G. Lees. Assessing the classification accuracy of multsource remote sensing data. *Remote Sensing of Enviroment*, 47:368–368, 1994.

[11] D. Geman and S. Geman. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6):721–741, Nov. 1984.

[12] A. D. Gordon. *Classification*. Chapman & Hall / CRC, 2 edition, 1999.

[13] Q. Jackson and D. A. Landgrebe. Adaptive Bayesian contextual classification based on Markov random fields. *IEEE Transactions on Geoscience and Remote Sensing*, 40(11):2454–2463, 2002.

[14] W. J. Krzanowski and F. H. Marriott. *Multivariate analysis: classification*. Arnold, 1995.

[15] J. R. Landis and G. C. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33:159–174, 1977.

[16] J. A. Matthews, E. M. Bridges, C. J. Caseldine, A. J. Luckman, G. Owen, A. H. Perry, R. A. Shakesby, R. P. D. Walsh, R. J. Whittaker, and K. J. Willis, editors. *The Encyclopaedic Dictionary of Environmental Change*. Arnold, London, 2001.

[17] D. K. Pickard. Inference for discrete Markov fields: The simplest nontrivial case. *Journal of the American Statistical Association*, 82(1):90–96, 1987.

[18] J. A. Richards and X. Jia. *Remote Sensing Digital Image Analysis: An Introduction*. Springer-Verlag, New York, 3 edition, 1999.

A. Parameters for each simulation situation

A.1. Situation P1

The parameters are those reported in [18, p. 188], obtained from an image with water, fire burn, vegetation and urban areas:

$$\boldsymbol{\mu}_1 = \begin{pmatrix} 44.27 \\ 28.82 \\ 22.77 \\ 13.89 \end{pmatrix}, \quad \boldsymbol{\mu}_2 = \begin{pmatrix} 42.85 \\ 35.02 \\ 35.96 \\ 29.04 \end{pmatrix},$$

$$\boldsymbol{\mu}_3 = \begin{pmatrix} 40.46 \\ 30.92 \\ 57.50 \\ 57.68 \end{pmatrix}, \quad \boldsymbol{\mu}_4 = \begin{pmatrix} 63.14 \\ 60.44 \\ 81.84 \\ 72.25 \end{pmatrix},$$

$$\boldsymbol{M}_1 = \begin{pmatrix} 14.36 & 9.55 & 4.49 & 1.19 \\ 9.55 & 10.51 & 3.71 & 1.11 \\ 4.49 & 3.71 & 6.95 & 4.05 \\ 1.19 & 1.11 & 4.05 & 7.65 \end{pmatrix},$$

$$\boldsymbol{M}_2 = \begin{pmatrix} 9.38 & 10.51 & 12.30 & 11.00 \\ 10.51 & 20.29 & 22.10 & 20.62 \\ 12.30 & 22.10 & 32.68 & 27.78 \\ 11.00 & 20.62 & 27.78 & 30.23 \end{pmatrix},$$

$$\boldsymbol{M}_3 = \begin{pmatrix} 5.56 & 3.91 & 2.04 & 1.43 \\ 3.91 & 7.46 & 1.96 & 0.56 \\ 2.04 & 1.96 & 19.75 & 19.71 \\ 1.43 & 0.56 & 19.71 & 29.27 \end{pmatrix},$$

$$\boldsymbol{M}_4 = \begin{pmatrix} 43.58 & 46.42 & 7.99 & -14.86 \\ 46.42 & 60.57 & 17.38 & -9.09 \\ 7.99 & 17.38 & 67.41 & 67.57 \\ -14.86 & -9.09 & 67.57 & 94.27 \end{pmatrix}.$$

A.2. Situation P2

Three bands and six classes; three of them with low mean values and the remaining three with high mean values, same covariance matrices for classes with close mean values:

$$\boldsymbol{\mu}_1 = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \quad \boldsymbol{\mu}_2 = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \quad \boldsymbol{\mu}_3 = \begin{pmatrix} 2 \\ 2 \\ 2 \end{pmatrix},$$

$$\boldsymbol{\mu}_4 = \begin{pmatrix} 125 \\ 125 \\ 125 \end{pmatrix}, \quad \boldsymbol{\mu}_5 = \begin{pmatrix} 142 \\ 142 \\ 142 \end{pmatrix}, \quad \boldsymbol{\mu}_6 = \begin{pmatrix} 234 \\ 234 \\ 234 \end{pmatrix},$$

$$\boldsymbol{M}_1 = \boldsymbol{M}_2 = \boldsymbol{M}_3 = \begin{pmatrix} 0.0100 & 0.0030 & 0.0009 \\ 0.0030 & 0.0100 & 0.0030 \\ 0.0009 & 0.0030 & 0.0100 \end{pmatrix},$$

$$\boldsymbol{M}_4 = \boldsymbol{M}_5 = \boldsymbol{M}_6 = \begin{pmatrix} 25.00 & 7.50 & 2.25 \\ 7.50 & 25.00 & 7.50 \\ 2.25 & 7.50 & 25.00 \end{pmatrix}.$$

A.3. Situation P3

Four classes and four bands; the classes have equal mean vectors and covariances, being differentiated by the variances only:

$$\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \boldsymbol{\mu}_3 = \boldsymbol{\mu}_4 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix},$$

$$\boldsymbol{M}_1 = \begin{pmatrix} 1.0000 & 0.3000 & 0.0900 & 0.0081 \\ 0.3000 & 1.0000 & 0.3000 & 0.0900 \\ 0.0900 & 0.3000 & 1.0000 & 0.3000 \\ 0.0081 & 0.0900 & 0.3000 & 1.0000 \end{pmatrix},$$

$$\boldsymbol{M}_2 = \begin{pmatrix} 2.0000 & 0.3000 & 0.0900 & 0.0081 \\ 0.3000 & 2.0000 & 0.3000 & 0.0900 \\ 0.0900 & 0.3000 & 2.0000 & 0.3000 \\ 0.0081 & 0.0900 & 0.3000 & 2.0000 \end{pmatrix},$$

$$\boldsymbol{M}_3 = \begin{pmatrix} 4.0000 & 0.3000 & 0.0900 & 0.0081 \\ 0.3000 & 4.0000 & 0.3000 & 0.0900 \\ 0.0900 & 0.3000 & 4.0000 & 0.3000 \\ 0.0081 & 0.0900 & 0.3000 & 4.0000 \end{pmatrix},$$

$$\boldsymbol{M}_4 = \begin{pmatrix} 8.0000 & 0.3000 & 0.0900 & 0.0081 \\ 0.3000 & 8.0000 & 0.3000 & 0.0900 \\ 0.0900 & 0.3000 & 8.0000 & 0.3000 \\ 0.0081 & 0.0900 & 0.3000 & 8.0000 \end{pmatrix}.$$

A.4. Situation P4

Six classes and three bands; the classes have equal zero mean vectors and proportional covariance matrices:

$$\boldsymbol{M}_1 = \begin{pmatrix} 1.00 & 0.30 & 0.09 \\ 0.30 & 1.00 & 0.30 \\ 0.09 & 0.30 & 1.00 \end{pmatrix}, \boldsymbol{M}_j = j^2 \boldsymbol{M}_1, 2 \leq j \leq 6.$$

B. Computational information

Developments were made in the IDL platform (`www.rsinc.com`) and incorporated into ENVI, an image processing platform developed in IDL. Plots were produced in R (`www.r-project.org`).