INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

A binary tree-structured MRF model for multispectral satellite image segmentation

Giuseppe Scarpa — Giovanni Poggi — Josiane Zerubia

N° 5062

December 2003

THÈME 3

Rapport de recherche

ISSN 0249-6399 ISRN INRIA/RR--5062--FR+ENG

A binary tree-structured MRF model for multispectral satellite image segmentation

Giuseppe Scarpa* , Giovanni Poggi * , Josiane Zerubia

Thème 3 — Interaction homme-machine,
images, données, connaissances
Projet Ariana

Rapport de recherche n° 5062 — December 2003 — 49 pages

Abstract: In this work we detail a tree-structured MRF (TS-MRF) prior model useful for segmentation of multispectral satellite images. This model allows a hierarchical representation of a 2-D field by the use of a sequence of binary MRFs, each corresponding to a node in the tree. In order to get good performances, one can fit the intrinsic structure of the data to the TS-MRF model, thereby defining a multi-parameter, flexible, MRF. Although a global MRF model is defined on the whole tree, optimization as well estimation can be carried out by working on a single node at a time, from the root down to the leaves, with a significant reduction in complexity. Indeed the overall algorithm is proved experimentally to be much faster than a comparable algorithm based on a conventional Ising MRF model, especially when the number of bands becomes very large. Thanks to the sequential optimization procedure, this model also addresses the cluster validation problem of unsupervised segmentation, through the use of a stopping condition local to each node. Experiments on a SPOT image of the Lannion Bay, a ground-truth of which is available, prove the superior performance of the algorithm w.r.t. some other MRF based algorithms for supervised segmentation, as well as w.r.t. some variational methods.

Key-words: Image classification, object-oriented segmentation, Bayesian estimation, hierarchical model, Markov random fields.

* Dipartimento di Ingegneria Elettronica e delle Telecomunicazioni – Universitá di Napoli "Federico II" – via Claudio 21, 80125, Napoli (Italy).

Unité de recherche INRIA Sophia Antipolis
2004, route des Lucioles, BP 93, 06902 Sophia Antipolis Cedex (France)
Téléphone : +33 4 92 38 77 77 — Télécopie : +33 4 92 38 77 65

Modèle MRF structuré à arbre binaire pour la segmentation d'images satellitaires multispectrales

Résumé : Dans ce rapport, nous proposons un modèle markovien *a priori* structuré à arbre binaire (le TS-MRF) pour la segmentation d'images satellitaires multispectrales. Ce modèle permet de représenter un champ bidimensionnel par une séquence de champs de Markov binaires, chacun correspondant à un noeud de l'arbre. Pour avoir une bonne classification, on peut adapter le modèle TS-MRF à la structure intrinsèque des données, en définissant un MRF, à plusieurs paramètres, très flexible. Bien que l'on définisse le modèle global sur tout l'arbre, l'optimisation et l'estimation peuvent être poursuivis en considérant un noeud à la fois, à partir de la racine jusqu'aux feuilles, avec une réduction significative de la complexité. En effet, on a montré expérimentalement que l'algorithme global est beaucoup plus rapide qu'un algorithme conventionnel fondé sur le modèle markovien d'Ising, en particulier quand le nombre des bandes spectrales est très grand. Grâce à la procédure d'optimisation séquentielle, ce modèle permet aussi de déterminer le nombre des classes présentes dans l'image satellitaire, dans le cadre d'une classification non supervisée, à travers une condition d'arrêt définie localement pour chaque noeud. Nous avons effectué des expériences sur une image SPOT de la baie de Lannion, pour laquelle nous disposons d'une vérité terrain, et nous avons trouvé que le modèle proposé fournit de meilleurs résultats que certains autres modèles de Markov et que d'autres méthodes variationnelles.

Mots-clés : Classification d'image, segmentation orientée aux objets, estimation Bayesienne, modèle hiérarchique, champs de Markov.

Contents

1 Introduction **4**
- 1.1 Multispectral image segmentation . . . 4
- 1.2 Statistical approach . . . 4
- 1.3 Discussion on the proposed MRF model . . . 6

2 MRF models for image segmentation **8**
- 2.1 MRF definition . . . 8
- 2.2 The Ising model . . . 8
- 2.3 Optimization . . . 9
- 2.4 Estimation . . . 9

3 Tree-structured MRF model **11**
- 3.1 Theoretical tree-structured MRF . . . 11
- 3.2 Recursive properties of TS-MRFs and optimization . . . 13
- 3.3 Model extension . . . 15
- 3.4 Adaptive Ising model . . . 16

4 Supervised segmentation of SPOT multispectral images **18**
- 4.1 The posterior model . . . 18
- 4.2 SPOT image of Lannion Bay . . . 19
- 4.3 Accuracy assessment method . . . 21
- 4.4 Reference methods . . . 21
- 4.5 TS-MRF implementations . . . 22
- 4.6 Experimental results . . . 23

5 Unsupervised image segmentation **36**
- 5.1 The unsupervised TS-MRF algorithm . . . 36
- 5.2 A split-and-merge solution . . . 39
- 5.3 Unconnected regions splitting . . . 40
- 5.4 Computational complexity . . . 42

6 Conclusions **44**

1 Introduction

1.1 Multispectral image segmentation

Segmentation is a low-level processing useful for many high-level algorithms and applications in the context of remote sensing, medical imaging, video coding and so on. Its goal is to partition the image in homogeneous regions according to a given set of features of interest. A notable example of the use of segmentation in the remote sensing field is classification, a task whose goal is to label each site of the scene as belonging to one of several known classes [24, 29]. From an algorithmic point of view, segmentation and classification can often be considered as the same task, *i.e.* the output of the former is directly used as a classification map, although the latter admits an higher level interpretation. This is the case, for example, when the spectral responses of the classes to be detected are known a priori, and such information is taken into account in the segmentation algorithm, which is necessarily parametric. In this case, the segmentation is said to be *supervised*, since several parameters, as well as the number of classes, are known a priori, avoiding the need of estimating them. On the contrary, when no prior information on the classes is available, the segmentation is said to be *unsupervised* and, if one resorts to the same parametric algorithms, it becomes much more complex, involving the estimation of both the number of classes (*cluster validation problem*) [1, 13] and the corresponding features, like mean and covariance of the classes [10, 35]. Indeed, when no information is given, the very concept of a class or segment is not obviously defined and, as a consequence, cluster identification is an ill-posed problem. Eventually, to regularize and solve this problem, several constraints, based on spatial, spectral or entropy properties, must be defined.

In this work, both the supervised and unsupervised cases are considered. In the first case, the experiments are carried out on a SPOT satellite image of Lannion Bay acquired during the summer 1997, coupled with the ground-truth of two sample sets of the scene detected by a human operator on site: the former (learning set) used to estimate the class parameters, the latter (test set) to evaluate the algorithm performance in terms of a correct classification percentage.

Unsupervised segmentation is also addressed thanks to the peculiar structure of the proposed model, which allows to define a simple cluster validation criterion, and experiments are provided on real-world hyperspectral data acquired by the GER (Geophysical Environmental Research) airborne sensor, which portrays an agricultural area in Germany near the river Rhein.

1.2 Statistical approach

In the statistical framework, the segmentation problem is approached by choosing an *ad hoc* probabilistic model to fit the data and the unknown segmentation map. In the basic formulation[1], multispectral image data are represented by a continuous vectorial 2-D field $y = \{y_s, s \in \mathcal{S}\}$, with $y_s \in \Re^B$, where s is a site of the lattice $\mathcal{S}$, and B is the number of the bands. The data are then assumed to be a realization of a random field Y whose probability distribution is $p(y)$[2]. Likewise,

[1] In this context the data are considered as raw, without any processing or transformation, and the segmentation is similarly represented as 2-D map although other points of view could be assumed (*e.g.*, contour set).

[2] Whenever unambiguous, we will indicate the probability law associated with X simply as $p(x)$, to be meant as either a density or a distribution function depending on the case.

the unknown segmentation map $x = \{x_s, s \in \mathcal{S}\} \in \Omega = \Lambda^{\mathcal{S}}$ (where $\Lambda = \{0, 1, 2, \ldots, K-1\}$ is the label set ordered only for notational convenience, and K is the number of the classes) is a realization of a discrete scalar 2-D random field X with distribution $p(x)$.

Given the probabilistic model, one can resort to a suitable estimation criterion (MAP, MMSE, ML, MPM, etc.) to minimize the mean of a selected cost function. One of the most popular is the *Maximum A Posteriori* (MAP) criterion whose related cost function is 0 if no errors occur, and 1 otherwise, irrespective of the number of errors. In the context of classification it is more convenient, to use a cost function which counts the number of misclassified pixels. Minimizing such a function leads to the *Maximum Posterior Marginal* (MPM) criterion which, however, is very hard to implement because of the complexity involved in computing the posterior marginal, unless the model has a specific form [6]. For this reason, the MAP criterion is much more frequently used in image segmentation. The MAP estimator is defined as

$$\widehat{x}_{MAP} = \arg\max_x p(x|y) = \arg\max_x p(y|x)p(x) \tag{1}$$

where the second equality comes from the Bayes formula once deleted the irrelevant term $p(y)$. When no a priori model $p(x)$ is assumed, which is equivalent to have a uniform a priori distribution, the maximization involves only $p(y|x)$, leading to the well known *Maximum Likelihood* (ML) estimator. In this case, the likelihood term $p(y|x)$ is usually assumed to be spatially independent, meaning that each site is independent from each other, with a Gaussian local conditional density whose parameters are class-dependent, that is

$$p(y|x) = \prod_{s \in \mathcal{S}} p(y_s|x_s) \tag{2}$$

and

$$p(y_s|x_s = k) = \frac{1}{(2\pi)^{B/2}|\Sigma_k|^{1/2}} \exp[-\frac{1}{2}(y_s - \mu_k)^T \Sigma_k^{-1}(y_s - \mu_k)] \tag{3}$$

where μ_k and Σ_k are the mean and the covariance matrix of class k respectively.

Under the mentioned hypotheses, thanks to the factorization, optimization can be pursued separately for each x_s, considerably reducing the computational burden. However, when the data are noisy, ML segmentation proves often unsatisfactory, having neglected any helpful contextual information, like the spatial correlation. To obtain acceptable results, one cannot rely solely on the observed data, but must take advantage of all available prior information about the image or class of images under analysis.

The Markov random field (MRF) modelling [2, 18, 25, 34] is a relatively simple, yet effective, tool to encompass prior knowledge in the segmentation process, and in fact the interest on MRFs has been steadily growing in recent years. When image segmentation is formulated as a Bayesian estimation problem, all prior information available on the image X to be segmented, must be contained in its probability distribution $p(x)$. By modelling the segmentation map as a MRF, *i.e.*, assuming that each given pixel X_s depends statistically on the rest of the image only through a selected group of neighbors $X_{\eta(s)}$, one simplifies the difficult problem of assigning a prior: one needs only specify the local characteristics of the image $p(x_s|x_{\eta(s)})$. What is more important, local dependencies can be conveniently expressed through the definition of suitable potential functions in a Gibbs distribution.

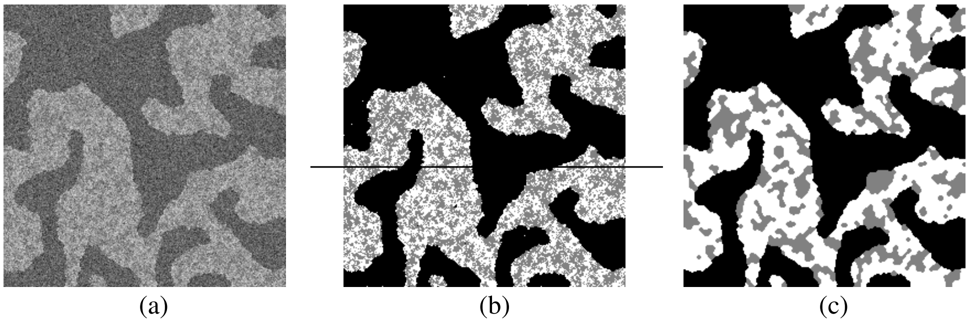


Figure 1: Synthetic experiment: (a) noisy data; (b) hidden segmentation map; (c) segmentation by Ising model.

1.3 Discussion on the proposed MRF model

In the MRF framework, a number of problems remain open, the most important being:

1. how to define a MRF (through its potentials) that is able to take into account prior information while remaining mathematically and numerically manageable;
2. how to set/estimate the numerical parameters of such an MRF;
3. how to solve the MAP estimation problem with reasonable computational complexity.

The first problem is certainly the most intriguing, as it amounts to defining an abstract structure of the image that fits well the observed data. One could be tempted to define sophisticated models, in order to capture the complex nature of image dependencies. However, model definition cannot overlook the estimation problems (2) and (3). In fact, by increasing the model complexity one ends up with a large number of parameters that cannot be reliably estimated; and even neglecting this problem, the subsequent optimization task could be so computationally demanding as to forbid the use of reliable procedures, leading to disappointing results. Indeed, computational complexity remains a major weakness of the MAP/MRF approach, and in developing a real-world MRF-based segmentation algorithm all efforts should be made to keep it under control, without sacrificing fidelity of description.

As is well known, one of the main problems in the image representation is to take into account the spatial variation of the statistics. Let's consider for example the image in Fig.1, where the data (a) were generated adding Gaussian noise to the synthetic image (b) which is supposed to be the unknown segmentation map. As can be seen, following for example the traced line, the class random process is clearly non-stationary. The well known Ising model [2], as well as more recent hierarchical MRFs [3, 7, 21, 23, 26], do not take into account these spatial variation and, as a consequence, gives poor results in the presence of such "structured" data. This is clearly shown in the example of Fig.1(c) where a segmentation map is achieved with a second order Ising model.

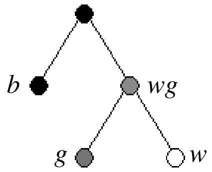


Figure 2: Tree structure of the data in Fig.1. *b*, black class; *g*, grey class; *w*, white class; *wg*, white-grey merging class.

This happens because in the prior model there is only one parameter to enforce regularity, whose value depends neither on the spatial position nor on the classes. As a consequence, one class is well detected while the other two are almost everywhere mixed because of an over-regularization.

The example shown, although not realistic, makes clear the need to use a structured model. In fact it is very easy to detect the intrinsic structure of the data (see Fig.2), which is well represented by a binary tree of classes. Indeed, most real images present a similar hierarchical structure even if it is not always so evident and easy to detect. The aim of the proposed model is just to find and represent the hidden hierarchy of the data by a *tree-structured* MRF (TS-MRF), which can be seen as a union of several binary MRFs each corresponding to a node in the tree. In such a structure, local parameters can be defined so as to correctly split the classes depending only on their statistical characteristics. In addition, thanks to its nested form, the TS-MRF model allows for recursive procedures of estimation and optimization, which help reducing computational complexity.

2 MRF models for image segmentation

2.1 MRF definition

As was said before, by using a MRF to model the prior probability $p(x)$ one obtains the a posteriori model which allows to exploit contextual information, and eventually to estimate the segmentation map according to a MAP criterion (1).

A random field X defined on a lattice $\mathcal{S}$ is said to be a MRF with respect to the neighborhood system $\eta = \{\eta(s), s \in \mathcal{S}\}$ if the Markovian property holds true for each site s, namely, X_s is conditionally independent of the rest of the image given the realization on a neighborhood $\eta(s)$ of s

$$p(x_s|x_{\mathcal{S}-s}) = p(x_s|x_{\eta(s)}). \tag{4}$$

Quite often, in order to limit complexity, only the 4 or 8 closest pixels (system η^1 and η^2, respectively) are included in a pixel's neighborhood (see Fig.3). Even so, the MRF model proves quite powerful because most of the dependencies can be captured through local interactions.

What makes MRFs especially attractive is the Hammersley-Clifford theorem [34] which proves that any positive MRF has a Gibbs probability law

$$p(x) = \frac{1}{Z}\exp[-U(x)] = \frac{1}{Z}\exp[-\sum_{c\in\mathcal{C}} V_c(x_c)] \tag{5}$$

where the $V_c(\cdot)$ functions are called potentials, $U(x)$ denotes the energy, and Z is just a normalizing constant. In (5) c indicates a clique of the image, which is a set of sites that are all neighbors of one another. Note that each potential V_c depends only on the values taken on the clique sites $x_c = \{x_s, s \in c\}$ and, therefore, accounts only for local interactions. As a consequence, local dependencies in X can be easily modeled by defining suitable potentials $V_c(\cdot)$, as shown for example in [34]. With the MRF-Gibbs equivalence, the model selection problem amounts to choosing a suitable neighborhood system and suitable potentials so as to express the desired spatial dependencies. A number of models have been proposed in the literature for various applications and we refer the reader to [25, 34] for a comprehensive review.

2.2 The Ising model

The Ising MRF model is one of the most widespread because of its very manageable form [2]. It can be defined both on η^1 and η^2, but in the latter case only two-site cliques are taken into account. For each clique a potential is defined as

$$V_c(x_c) = V_c(x_p, x_q) = \begin{cases} \beta & \text{if } x_p \neq x_q, \quad p,q \in c \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

Single-site cliques are not used because there is no reason to favor a label over another, and larger cliques are neglected to speed-up processing. Once given the potential functions, the global distribution $p(x)$ is completely defined, and the local characteristics $p(x_s|x_{\eta(s)})$ can be expressed [18] as

$$p(x_s = k|x_{\eta(s)}) \propto \exp[\beta N_k] \propto \exp[-\beta N_{\bar{k}}], \tag{7}$$

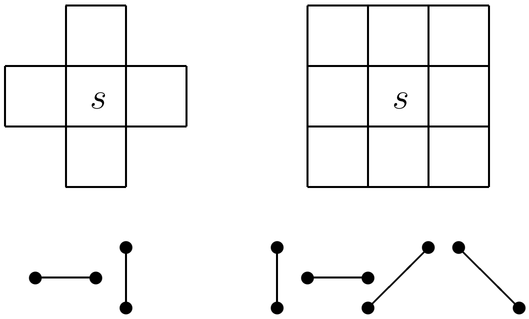


Figure 3: Pixel s and its neighbors in the η^1 (left) and η^2 (right) system. In the η^2 system diagonal two-site cliques can be used (neglecting more complex cliques for η^2 which are often dropped in practical models).

where N_k ($N_{\bar{k}}$) is the number of neighbors of s with label k (different from k).

With this model, the vector of parameters θ associated with the prior model $p(x)$ reduces to a single parameter $\beta > 0$, interpreted as an "edge-penalty". In fact, when $\beta = 0$ all realizations are equally likely, whereas larger values of β tend to penalize non-homogeneous cliques making smoother realizations more and more likely. Of course, β is not known a priori, and must be estimated together with x.

2.3 Optimization

Assuming that one has been able to select a satisfactory model for the prior distribution, the problem remains of maximizing $p(x)p(y|x)$ over x, which is when computational complexity becomes a dominant concern. It can be shown that $p(x)p(y|x)$ is itself a Gibbs distribution, and its maximization can be carried out via the simulated annealing (SA) technique [18], but only under conditions (a very slow cooling schedule) that make it unfeasible in practice. Using a faster schedule, on the other hand, compromises SA's optimality. Gradient-based algorithms tend to remain trapped in local optima as the objective function is in general non-convex. This is the case of the iterated conditional modes (ICM) algorithm [34], structurally similar to the SA but much faster, which is why it is often the algorithm of choice. Even resorting to the ICM, and hence renouncing global optimality, computational complexity remains the major bottleneck of MRF-based segmentation, and the problem only worsens when there is no *a-priori* knowledge on the image, namely in unsupervised segmentation.

2.4 Estimation

Quite often, indeed, a number of important parameters such as the number K of labels/classes in the image, the parameters of the likelihood term $p(y|x)$, and the parameters of the Gibbs prior $p(x)$, are not known, and must be estimated from the data together with the segmentation $\widehat{x}$ itself. The single most critical parameter is by far the number of classes K, since it influences heavily all other aspects of segmentation.

The problem of determining the number of classes in a data set, or cluster validation problem, has received a great deal of attention in the literature [13], with mixed and inconclusive results. As a matter of fact, in a real-world image, the number of different segments that can be identified varies wildly according to the user's point of view. In a remote-sensing image, for example, a single segment labeled as "urban area" in one application, could be further partitioned into smaller segments in another application. In the absence of prior information on the application, both solutions are equally reasonable, and both should be preserved to let a human interpreter have a final say. Although some efficient strategies have been proposed to address the cluster validation problem, especially in the context of split-and-merge approaches, this is still one of the main reasons for the increase in complexity going from supervised to unsupervised segmentation.

Another reason is the need to estimate, together with the segmentation, the parameters of the involved distributions, collectively represented by a random vector Θ

$$(\widehat{x}, \widehat{\theta}) = \arg\max_{x,\theta} p(x, y|\theta) \tag{8}$$

Since exact joint optimization is computationally intractable, a two-step procedure is often used. First, the model parameters are estimated from the observed data, following for example an ML approach, then the MAP segmentation is carried out in a second step using the estimated parameter values. A number of techniques can be used to perform the ML parameter estimation, such as the EM algorithm and its numerous variants, or the similar but more general ICE [31]. Except for some simple cases, however, these algorithms do not have an analytical closed form, and are quite computationally expensive. For this reason, we will resort to the suboptimal, but much simpler, alternating marginal optimization ($\widehat{x}$ and $\widehat{\theta}$ are alternately optimized given each other) which can be viewed as an approximation of the two step EM-approach, and has been observed to provide comparable results in various practical situations [27].

3 Tree-structured MRF model

The algorithm discussed here, and proposed in its basic form in [28], is based on the recursive binary segmentation of image regions. As discussed in Section 1.3, commonly used MRF models impose some constraints on the clique's potential functions in order to reduce the number of parameters involved in the model. Let us consider, in particular, the Ising model, where the potential function of a two-site clique depends only on the presence or absence of an edge (a class transition), without regard to involved classes. As a consequence one parameter is sufficient to define the model. If we remove such a constraint and use a different potential function for each different pair of classes, a much larger number of parameters becomes necessary, $\frac{1}{2}(K-1)K$ where K is the number of classes. In this generalized Ising model, the local characteristics will become

$$p(x_s = k|x_{\eta(s)}) = \frac{1}{Z}\exp[-\sum_{h\neq k}\beta_{kh}N_h], \tag{9}$$

where $\beta_{kh} = \beta_{hk}$, $k \neq h$, is the edge-penalty parameter for a transition $k-h$, and Z is again a normalizing constant. Accordingly, the clique potentials will be defined as

$$V_c(x_c) = V_c(x_p, x_q) = \begin{cases} \beta_{kh} & \text{if } x_p = k \neq x_q = h, \quad p, q \in c \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

Starting from this very general model, it is possible to reduce the number of parameters by taking into account some class properties. Looking at the example of Fig.1, it is easy to realize that the white (*w*) and gray (*g*) classes have the same relationship with the black class (*b*), that is, a *b-w* edge is statistically equivalent to a *b-g* edge. Therefore, only two independent parameters should be used instead of three. Such class properties can be easily represented by means of a hidden binary structure[3] (see Fig.2). In the example considered, the two relevant parameters are associated with the two internal nodes of the tree. The former, associated with the root, controls the edges *b-w* and *b-g* (which are assumed to be equivalent), while the latter, associated with the other internal node, controls the edges *w-g*. In other words, at the root level, where *b* is split from *w* and *g*, it does not matter if a site is labeled as *w* or *g*, but only if it falls into the macro-class *wg* (merging of *w* and *g*).

In general, given a binary structure with K terminal nodes, the number of internal nodes, and hence the number of parameters, will be $K-1$, rather than $\frac{1}{2}(K-1)K$. More in general, even considering non-isotropic models and/or more sophisticated cliques, one gets the same parameter reduction ratio ($K/2$) between a complete unconstrained model and the "tree-structured" dual one. Looking at the estimation problem, if the data can be well represented by this kind of structure, the information available to estimate its few parameters will increase, resulting in better estimates.

3.1 Theoretical tree-structured MRF

Let us first define a theoretical tree-structured MRF model, and later the model that has been actually implemented. To this end, let us consider a binary tree T, identified by its nodes and their mutual

[3] A more general tree-structure can also be defined, but we consider only binary trees here for the sake of simplicity.

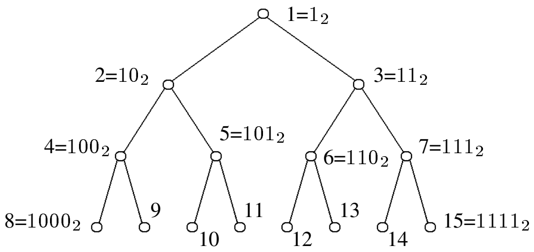


Figure 4: Tree indexing.

relationships. Except for the root, each node t has one parent $u(t)$, and each internal node has two children $l(t)$ and $r(t)$, with $u[l(t)] = u[r(t)] = t$. We also define $\widetilde{T} = \{t \in T : l(t) = r(t) = \emptyset\}$, the set of terminal nodes or leaves, and $\overline{T} = T - \widetilde{T}$, the set of internal nodes.

We will use integer numbers to index the nodes of the tree, as well as all items associated with them, so that root $= 1$, $l(t) = 2t$, $r(t) = 2t+1$ and $u(t) = \lfloor t/2 \rfloor$ (see Fig.4). Note that each terminal node corresponds to a class, while each internal node corresponds to both a merging class and an edge-penalty parameter. In order to define the model it is helpful to use the binary representation of the indexing integers. Let $\nu(t)$ be the function that converts a non negative integer $t \in \mathbb{N}$ to its corresponding variable-length binary code $c \in \mathcal{B}$, where all leading zeros are discarded (see the balanced tree of Fig.4), and let ℓ_t be the corresponding length. Let us also define the function $\Psi(a,b) : \mathcal{B} \times \mathcal{B} \rightarrow \mathcal{B}$ which returns the longest common prefix of a and b. It's easy to check that $\Psi(a,b)$ gives the nearest common ancestor node of a and b.

We now define a tree-structured MRF through its local characteristics, still expressed by (9), but with the additional $\frac{1}{2}K(K-3)+1$ constraints

$$\beta_{kh} = \beta_{pq} = \beta_t \tag{11}$$

for (k,h) and (p,q) such that

$$\Psi(\nu(k), \nu(h)) = \Psi(\nu(p), \nu(q)) = \nu(t), \tag{12}$$

with $k \neq h$, $p \neq q$, $(k,h) \neq (p,q)$.

Reorganizing the terms in (9) we can explicit the local characteristics with respect to the non-redundant parameter set $\{\beta_t\}_{t \in \overline{T}}$ as follows:

$$p(x_s = k | x_{\eta(s)}) = \frac{1}{Z} \exp[-\sum_{n=1}^{\ell_k - 1} \beta_{\nu^{-1}(k_1,\ldots,k_n)} N_{\nu^{-1}(k_1,\ldots,k_n,\overline{k}_{n+1})}], \tag{13}$$

with $\nu(k) = (k_1, \ldots, k_{\ell_k})$. Here, when h corresponds to an internal node, $x_s = h$ means that s belongs to one of the descendant classes of h, and N_h is the number of neighbours of s which belong to one of such classes. For example, with reference to Fig.2, we have

$$\begin{aligned} p(x_s = 10|x_{\eta(s)}) &= \frac{1}{Z}\exp[-\beta_1 N_3 - \beta_2 N_4 - \beta_5 N_{11}] \qquad (14)\\ &= \frac{1}{Z}\exp[-\beta_1(N_{12}+N_{13}+N_{14}+N_{15}) - \beta_2(N_8+N_9) - \beta_5 N_{11}] \end{aligned}$$

The clique potentials are expressed by

$$V_c(x_c) = V_c(x_p, x_q) = \begin{cases} \beta_{\nu^{-1}(\Psi(\nu(k),\nu(h)))} & \text{if } x_p = k \neq x_q = h, \quad p,q \in c \\ 0 & \text{otherwise} \end{cases} \qquad (15)$$

Now we define the function $\mathcal{N}_t(x)$ which gives the number of cliques in the map x with edge-penalty β_t. With this position, the joint probability of the TS-MRF becomes simply

$$p(x) = \frac{1}{Z}\exp[-\sum_{t\in\overline{T}} \beta_t \mathcal{N}_t(x)]. \qquad (16)$$

The complexity of this model could still seem prohibitive for a practical implementation because of the dimensionality of the parameter space, dependent on the number of classes, that makes very hard the optimization. However, thanks to the structural constraints of the model, a recursive optimization procedure can be used which, although sub-optimal, involves only one edge-penalty at a time.

3.2 Recursive properties of TS-MRFs and optimization

Let us consider for each node t of a tree T,

- a set of sites $\mathcal{S}^t \subseteq \mathcal{S}$, corresponding to a segment of the image (in particular $\mathcal{S}^{\mathrm{root}} = \mathcal{S}$);
- a binary random field $X^t = \{X_s^t : s \in \mathcal{S}^t\}$, with realization x^t where $x_s^t \in \{l(t), r(t)\}$.

Now we impose the additional constraint that the set of sites associated with any given node is obtained from the binary segmentation of the parent set of sites. More formally, for each internal node of the tree $t \in \overline{T}$

$$\begin{cases} \mathcal{S}^{l(t)} &= \{s \in \mathcal{S}^t : x_s^t = l(t)\} \\ \mathcal{S}^{r(t)} &= \{s \in \mathcal{S}^t : x_s^t = r(t)\} \end{cases} \qquad (17)$$

Therefore, the tree-structured MRF X is completely given by the set of binary fields $\{X^t\}_{t\in\overline{T}}$ and *vice-versa*, that is

$$X = \bigcup_{t\in\overline{T}} X^t. \qquad (18)$$

Let us define, now, $\omega(t) = \{h \in \overline{T} - \{t\} : \nu(h) \text{ is a prefix of } \nu(t)\}$, the set of the ancestor nodes of t, and $X^{\omega(t)} = \{X^t\}_{t\in\omega(t)}$, the set of the ancestor fields of t (of course, $\omega(1) = X^{\omega(1)} = \emptyset$).

Observe that, except for X^1, each field X^t depends on the ancestor fields $\{X^{\omega(t)}\}$, in particular, the very same domain of X^t is fixed once the ancestor fields are specified. On the other hand, given a realization $x \equiv \{x^k\}_{k\in\overline{T}}$, the number $\mathcal{N}_t = \mathcal{N}_t(x)$ of cliques with edge-penalty β_t depends only on x^t and, for the above considerations, on $x^{\omega(t)}$, while it is independent of other component binary fields. As a consequence, the joint probability of the overall field (16) becomes

$$
\begin{aligned}
p(x) &= \frac{1}{Z}\exp[-\sum_{t\in\overline{T}}\beta_t\mathcal{N}_t(x^t, x^{\omega(t)})] && (19)\\
&= \prod_{t\in\overline{T}}\frac{1}{Z_t}\exp[-\beta_t\mathcal{N}_t(x^t, x^{\omega(t)})]. && (20)
\end{aligned}
$$

It is also easy to prove that, for each node in the tree, given X^t and $X^{\omega(t)}$, the set of fields which lie on the left sub-tree stemming from t is independent from the set of fields which lie on the right sub-tree. As an example, for the structure in Fig.4 we can write

$$
\begin{aligned}
p(x^5, x^4, x^2|x^7, x^6, x^3, x^1) &= \frac{p(x)}{p(x^7, x^6, x^3, x^1)} && (21)\\
&= \frac{\frac{1}{Z}\exp[-\sum_{t=1}^{7}\beta_t\mathcal{N}_t]}{\sum_{x^5,x^4,x^2}\frac{1}{Z}\exp[-\sum_{t=1}^{7}\beta_t\mathcal{N}_t]} && (22)\\
&= \frac{\frac{1}{Z}\exp[-\beta_2\mathcal{N}_2-\beta_4\mathcal{N}_4-\beta_5\mathcal{N}_5]}{\sum_{x^5,x^4,x^2}\frac{1}{Z}\exp[-\beta_2\mathcal{N}_2-\beta_4\mathcal{N}_4-\beta_5\mathcal{N}_5]} && (23)\\
&= \frac{1}{Z(x^1)}\exp[-\beta_2\mathcal{N}_2-\beta_4\mathcal{N}_4-\beta_5\mathcal{N}_5] && (24)\\
&= p(x^5, x^4, x^2|x^1), && (25)
\end{aligned}
$$

which proves the independence. In a similar way, it can be proved that $p(x^4|x^5, x^2, x^1) = p(x^4|x^2, x^1)$ and so on. More in general, thanks to the above property, by a recursive use of the Bayes theorem we have

$$p(x) = \prod_{t\in\overline{T}} p(x^t|x^{\omega(t)}). \qquad (26)$$

Note also that, given the ancestor field X^1, the field built on the sub-tree with root in $t = 2$, (X^2, X^4, X^5), is still a TS-MRF (24); this property holds for each internal node t as well. As a consequence, given $X^{\omega(t)}$, the terminal binary fields X^t (associated with terminal splits) are Ising MRFs, that is

$$p(x^t|x^{\omega(t)}) = \frac{1}{Z(x^{\omega(t)})}\exp[-\beta_t\mathcal{N}_t], \qquad (27)$$

This property does not hold for non-terminal binary fields, because, in this case, the partition function Z is itself a function of x^t. For example we have

$$\begin{aligned} p(x^2|x^{\omega(2)}) &= \sum_{x^5,x^4} p(x^5,x^4,x^2|x^1) && (28) \\ &= \frac{1}{Z(x^1)}\exp[-\beta_2\mathcal{N}_2]\sum_{x^5,x^4}\exp[-\beta_4\mathcal{N}_4-\beta_5\mathcal{N}_5] && (29) \\ &= \frac{1}{Z(x^1)Z(x^1,x^2)}\exp[-\beta_2\mathcal{N}_2]. && (30) \end{aligned}$$

In other words, not all the terms of (26) are Ising distributions, as one could believe for the similarity between (20) and (26). Nonetheless, in order to find a MAP estimate of a segmentation with TS-MRF prior probability, one can recursively maximize the terms in (20), together with the likelihood parts, starting from the root and descending the tree until all leaves are reached. Each term depends only on a binary field X^t once its ancestor fields $x^{\omega(t)}$ are given and, also, it does have an Ising form. As a consequence, each one can be maximized, just like with an ordinary Ising MRF, by using simulated annealing, ICM, etc. Note, again, that in the step corresponding to node t, only the parameter β_t must be estimated, and that $\mathcal{N}_t$ is a sufficient statistic for β_t. Therefore, when the prior parameters are unknown, estimation-maximization procedures can be used again following a recursive schedule.

Finally, we underline that each binary field X^t, except for the root field, makes sense only once the realization $x^{\omega(t)}$ of its ancestor fields are given, since it is defined on an irregular (that is, non-rectangular) lattice whose shape is a result of $x^{\omega(t)}$.

3.3 Model extension

To build the TS-MRF model we started from the Ising model, generalized it, and finally introduced some structural constraints. More generally, one can start from any MRF model whose energy function depends linearly on the parameter vector, so that its distribution can be expressed as

$$p(x) = \frac{1}{Z}\exp[-\sum_{t\in\overline{T}}\langle\beta_t,\mathcal{N}_t\rangle]. \qquad (31)$$

where $\langle\cdot,\cdot\rangle$ indicates the scalar product, β_t is the parameter vector for node t, and $\mathcal{N}_t = \mathcal{N}_t(x^t, x^{\omega(t)})$ is a vector function which depends only on x^t and, of course, its ancestor fields $x^{\omega(t)}$.

In order to get such an expression, each clique potential must be a function of two labels (which are related in same way by the fixed hierarchical structure). Therefore, if cliques with more than two sites are considered, such a condition cannot be satisfied. However, if all clique potentials are linear combinations of terms that depend each on a couple of labels, then the model is still valid for tree-structured extension.

Given the above consideration, we can point out that the TS-MRF model is quite general and flexible, allowing to use high-order cliques (e.g., the Chien model [11, 12]), for a specific couple of classes, while all the other couples are described by much simpler cliques. This could be useful, for example, when two classes are highly correlated and very hard to divide. In such a case only an high-order interaction MRF model can provide good results, but such a model would be computationally intractable when defined on the whole image and for all the classes into it. On the contrary, we can use such model only for a terminal split of an *ad hoc* TS-MRF (the split of the two highly correlated classes). By doing so, the recursive optimization of the TS-MRF will have a reduced computational burden on almost all the tree except for the limited region where the complex model is used.

3.4 Adaptive Ising model

We present now a MRF model, referred here as "adaptive" Ising, which proved to be useful as basic component for a tree-structured model, especially in the presence of low resolution data. Both in the TS-MRF case and the flat-MRF case, during the segmentation process several region boundaries are "fragmented", namely, there are many boundary pixels which are associated with none of the adjacent regions but have a different label attached. Quite often this phenomenon does not correspond to a ground truth, namely, there is no actual "third" region between the two neighboring regions. Instead, it is likely due to the finite resolution of the sensor which, in boundary cells, happen to integrate contributions from several land covers, creating a spectral response that is quite different from those of the adjacent regions. Therefore, in boundary cells, observed data are not very reliable.

In a binary split, the pixel is attributed to one of the two adjacent regions, but when this region is further split, very likely this "uncertain" pixel will be erroneously classified. Fig.5 clarifies this point by showing how fragmentation arises in a small region selected from a larger image. The first split separates the bright region on the left from the dark region on the right: some boundary pixels with intermediate characteristics are classified as dark. The second split operates on the dark region and further divides it: at this point, because of their mixed nature, some of the former boundary pixels are now attributed to the wrong subregion. Obviously, this makes much less sense than choosing either one of the neighboring regions and should be avoided unless the data give a clear indication. In the selected example, the segmentation map shows that along the boundary between the two regions there are pixels wrongly classified as belonging to other classes.

It is clear that the data by their very nature tend to cause such wrong segmentations near region boundaries, and in fact this phenomenon arises just as well when flat MRF models are used. In addition, prior information does not help correctly resolving such ties, because the mixed neighborhood gives fuzzy indications. However, when a tree-structured segmentation is used, a simple fix is available to make better use of contextual information [15]. In such a solution, the weight of the context versus the likelihood (data attachment) varies spatially, by varying the edge-penalty β which controls the prior distribution. In particular, a parameter β, estimated with classical approaches, is used almost everywhere except on the cliques which are neighbours of boundaries coming from ancestor splits, where a larger parameter is used according to a reasonable heuristic (see [15] for details).

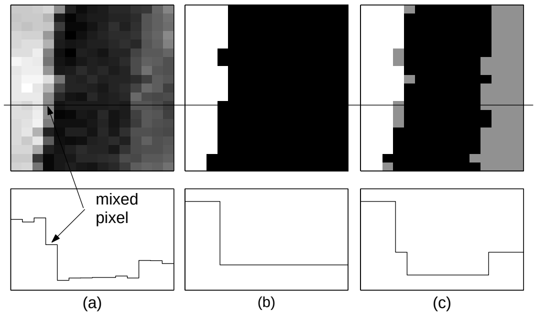


Figure 5: Boundary pixels between two regions often have mixed features (a); they can belong to either one of the neighboring regions (b), but in subsequent splits can be erroneously associated with a “third” class (c).

4 Supervised segmentation of SPOT multispectral images

4.1 The posterior model

The previous section deals with the prior probabilistic model useful for contextual image segmentation, and in particular we introduced the TS-MRF model and discussed its properties. Now we look at the joint distribution $p(x, y)$ to maximize in order to get the MAP estimate of the segmentation (1), thus including the likelihood term $p(y|x)$. It is easy to prove that this is still an MRF distribution, approximatively of the same kind of $p(x)$. In particular, they differ only for the introduction of unary cliques in the posterior model, which takes into account the likelihood probability. In fact, using (2) and (3) we get

$$\begin{aligned} p(y|x) &= \prod_{s\in\mathcal{S}} \frac{1}{(2\pi)^{B/2}|\Sigma_{x_s}|^{1/2}} \exp[-\frac{1}{2}(y_s-\mu_{x_s})^T\Sigma_{x_s}^{-1}(y_s-\mu_{x_s})] && (32)\\ &= \frac{1}{\widetilde{Z}} \exp[-\sum_{s\in\mathcal{S}} \frac{1}{2}\ln|\Sigma_{x_s}| + \frac{1}{2}(y_s-\mu_{x_s})^T\Sigma_{x_s}^{-1}(y_s-\mu_{x_s})] && (33)\\ &= \frac{1}{\widetilde{Z}} \exp[-\sum_{s\in\mathcal{S}} \widetilde{V}_s(x_s)], && (34)\end{aligned}$$

where we defined the unary clique potential functions[4]

$$\widetilde{V}_s(x_s = k) = \frac{1}{2}\ln|\Sigma_k| + \frac{1}{2}(y_s-\mu_k)^T\Sigma_k^{-1}(y_s-\mu_k). \quad (35)$$

Therefore, for each MRF prior $p(x)$, with clique set $\mathcal{C}$ and potential function $V_c(\cdot)$, we can express the posterior distribution as

$$\begin{aligned} p(x|y) &= \frac{p(y|x)p(x)}{p(y)} && (36)\\ &= \frac{1}{p(y)\widetilde{Z}Z} \exp[-\sum_{s\in\mathcal{S}} \widetilde{V}_s(x_s) - \sum_{c\in\mathcal{C}} V_c(x_c)] && (37)\\ &= \frac{1}{Z'} \exp[-\sum_{c\in\mathcal{C}'} V'_c(x_c)], && (38)\end{aligned}$$

where $\mathcal{C}' = \mathcal{C} \bigcup \mathcal{S}$ and $V'_c(\cdot)$ are potential functions equal to $V_c(\cdot)$ or $\widetilde{V}_c(\cdot)$ if c is a clique of $\mathcal{C}$ or a single pixel respectively.

Let us now consider the TS-MRF case, and in particular the binary field X^t associated with node t. Let $Y^t = \{Y_s : s \in \mathcal{S}, X_s = t\} \subseteq Y$ be the set of data whose labels belong to some descendant class of t, which is known given $x^{\omega(t)}$. As we said before, each X^t can be considered as a binary

[4] We neglected the dependence of $\widetilde{V}_s(\cdot)$ from y_s since the data y are given, *i.e.* constant in the problem.

Ising field in order to implement the recursive maximization procedure. Indeed, we have to consider a posterior distribution written as

$$p(x^t|x^{\omega(t)}, y^t) \propto \exp[-\beta_t \mathcal{N}_t] p(y^t|x^t, x^{\omega(t)}) = \exp[-\beta_t \mathcal{N}_t] \prod_s p(y_s^t|x_s^t, x_s^{\omega(t)}). \tag{39}$$

Here, the likelihood term $p(y^t|x^t, x^{\omega(t)})$ needs to be better defined. In fact, since the descendant fields of node t are unknown for the time being, we are only deciding, for each site, if it belongs to some of the left or right descendant classes, without exactly specifying which one. Therefore, we don't know which normal distributions to use to carry out the test.

To solve this problem, we propose two strategies. The first one, better described in section 5, and useful for unsupervised segmentation, makes use of an auxiliary normal distribution for each internal node corresponding to a "template" class seen as the merging of more classes. This is necessary because, in the unsupervised TS-MRF algorithm [17], neither the tree structure nor the classes are known in advance; these are provided automatically by the recursive tree growth procedure which is based on statistical tests, local to each node, that indicates whether the class must be further split or not. The second strategy, that we are going to describe here, is a more natural solution in the supervised case. Nonetheless, it can be easily generalized as well for use in unsupervised segmentation.

Let us consider the left, $l(t)$, and right, $r(t)$, children of an internal node t, and define $\gamma(k) = \{t \in \widetilde{T} : \nu(k) \text{ is a prefix of } \nu(t)\}$, the set of the descendant leaves of k. Now we can define the likelihood terms of (39) as

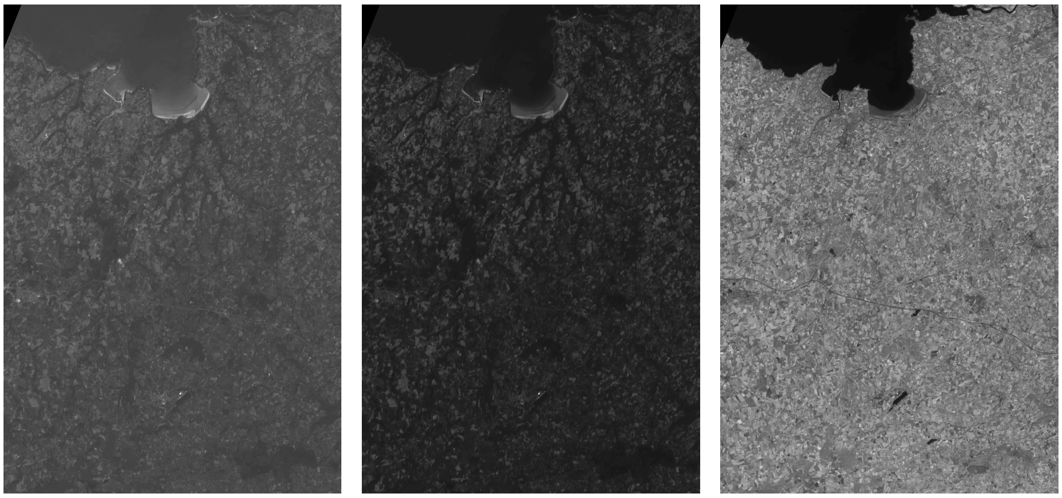
$$p(y_s^t|x_s^t, x_s^{\omega(t)}) = \max_{k \in \gamma(x_s^t)} p(y_s|x_s = k) \tag{40}$$

where $x_s^t \in \{l(t), r(t)\}$ and $p(y_s|x_s = k)$ are the normal densities given in (3). In other words, to decide if the current site should be assigned to the left or right node, the best two Gaussian distributions corresponding to "true" classes are considered, one being the most likely in $\gamma(l(t))$, the other in $\gamma(r(t))$. This way to proceed means that the tree-structure involves only the prior MRF model while no structural constraints are transferred on the likelihood term $p(y|x)$.

Note that the best fitting Gaussian chosen at this point is only a temporary choice, taken to well fit the data during this intermediate split, but further splits can change such decision based on newly available contextual information.

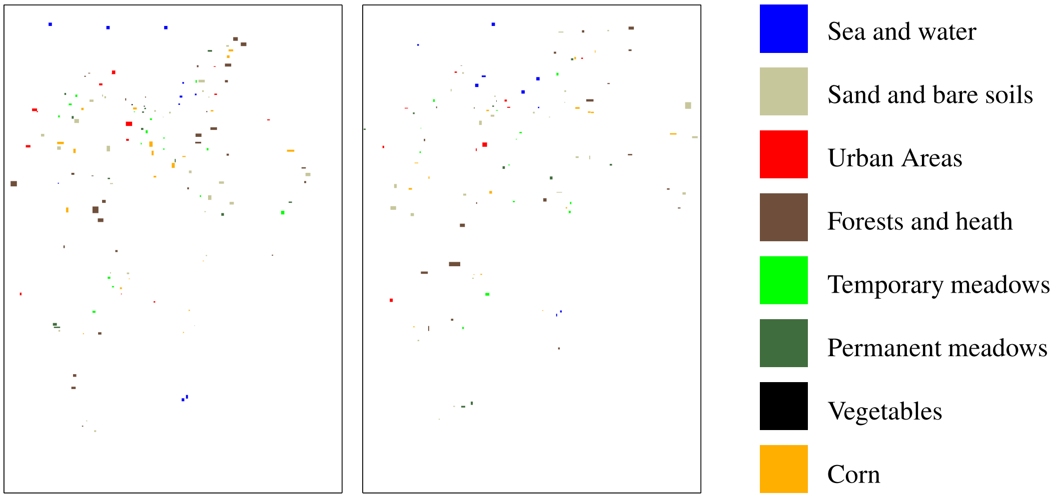
4.2 SPOT image of Lannion Bay

The supervised TS-MRF algorithm was applied to SPOT satellite images. The scene (Fig.6) is composed of three 1480×1024 images with different wavelengths in the visible spectrum and represents the Bay of Lannion in France in August 1997. The goal of this study was to determine the land cover of this area. So as to reach this aim, the geographers of the Costel laboratory (University of Rennes 2) built a list of eight classification categories: sea and water, sand and bare soil, urban areas, forests and heath, temporary meadows, permanent meadows, vegetables, corn.



Channel XS1 Channel XS2 Channel XS3

Figure 6: SPOT multispectral image of Lannion Bay (1480 x 1024) . August 15, 1997.



(a) Learning set (b) Test set (c) Classes

Figure 7: Ground-truth. Training set (a), validation set (b) and classes (c).

Thanks to both tests on the land and photointerpretation, they were also able to provide samples of these eight categories on the multispectral SPOT image of the scene. Some of them (learning test) were used to learn the mean spectral response and the inter-band covariance matrix of each category, so that we could perform supervised classifications, while the remaining samples (test set) were kept to assess the accuracy of the classifications (see Fig.7).

4.3 Accuracy assessment method

By the use of the test set, the accuracy of each classification method is assessed based on its confusion matrix. Recall that the entry of ith row and jth column of this matrix is the number of sample pixels from jth class that have been classified as belonging to the ith class.

Various indicators are derived from this matrix. First, two error assessments can be computed for each class: The user's accuracy of class i is defined as a_{ii}/a_{i+}, where a_{i+} is the ith row marginal (sum of row entries); conversely, the producer's accuracy of this class is defined as a_{ii}/a_{+i}, where a_{+i} is the ith column marginal.

Beside these two class-based parameters, three global quality indicators are also computed. The overall accuracy of the method defined as $\tau = \sum_i a_{ii}/N$, is the percentage of sample pixels that are well classified. Based on the confusion matrix, another indicator is the so-called Kappa parameter, which is defined as $\kappa = (N \sum_i a_{ii} - \sum_i a_{i+}a_{+i})/(N^2 - \sum_i a_{i+}a_{+i})$. Finally, it might be advantageous to normalize the confusion matrix with the iterative proportional fitting algorithm, such that all column and row marginals are equal to 1 [9]. The overall accuracy computed on this modified matrix is the normalized accuracy τ^{norm}.

4.4 Reference methods

Several researchers working on segmentation have used the SPOT data of Fig.6 and, as a consequence, we can benefit from the numerical results that they have obtained [5, 19, 32]. We collected the available information as belonging to two different experiments, one performed on the whole image (1480×1024) and another performed on a smaller sub-image of it (400×400). Some of the segmentation maps and numerical results obtained by such methods were not available and/or not published and were eventually neglected.

The reference methods are listed below:

- MD, minimum distance;
- ML, maximum likelihood;
- DA, discriminant analysis;
- ISING, Ising MRF optimized by ICM;
- H-MAP (H-MPM), hierarchical MRF model optimized by MAP (MPM) criteria;
- M1X and M2X, two variational models.

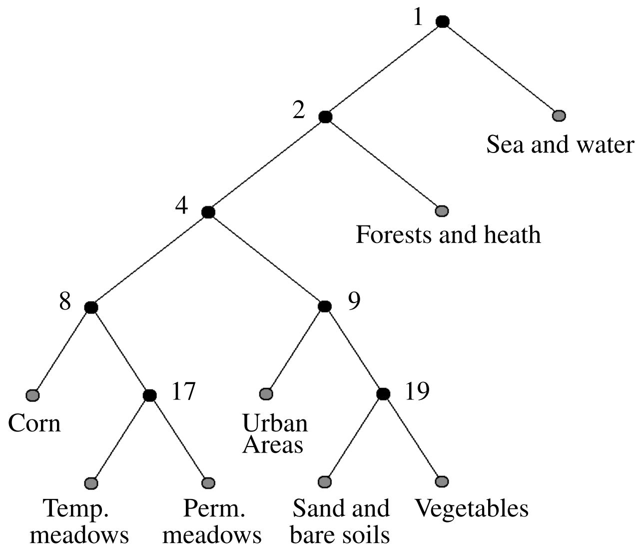


Figure 8: TS-MRF structure for the segmentation of the SPOT image.

MD, ML and DA are well known spectral segmentation techniques, while ISING is a MRF-based contextual segmentation algorithm. Such techniques were used as reference in the PhD thesis of Annabelle Chardin [5, 19], where a hierarchical MRF model was proposed providing two implementation: H-MAP and H-MPM. Later on, two variational methods, M1X and M2X, were proposed by C.Samson et al. [32] and compared with MV, ICM and H-MAP, experimenting on the small SPOT image.

4.5 TS-MRF implementations

All the algorithms cited above, as well as the TS-MRF versions used for the same experiment, are supervised, meaning that the number of classes and the associated parameters are estimated on the learning set. However, for the TS-MRF one has also to select one of the possible binary structures, whose leaves will correspond to the classes. Such a choice is crucial and, as said in the introduction, it has to reflect the "intrinsic" structure of the data, which means that the classes have to grouped together, hierarchically, taking into account their mutual spatial and spectral correlations. Although in the case of unsupervised TS-MRF a test parameter, named *split-gain*, helps to automatically solve this problem, in this case we defined the structure to use (see Fig.8) by inspection of the data.

We remark, also, that all the reference algorithms assume the three spectral bands of the SPOT image to be uncorrelated, namely, the covariance matrix for each class to be diagonal, and compute only the variances for each band. Therefore, in our experiments, we used both uncorrelated bands, to get a fair comparison, and correlated bands, to show the relative effectiveness of this second method.

In addition, we considered other two choices, one where only binary Ising MRFs are associated with each node of the tree, the other, called mixed solution, is similar to the previous one with the exception of the two nodes 9 and 19 whose corresponding fields are "adaptive" Ising. This last solution provides some improvements. Eventually, we implemented four TS-MRF versions which we will refer as

- TS.U, uncorrelated channels and Ising components,
- TS.U+, uncorrelated channels and mixed components,
- TS.C, correlated channels and Ising components,
- TS.C+, correlated channels and mixed components.

Finally, for each binary split we used an estimation-maximization schedule similar to that proposed in [22], but with the use of ICM instead of simulated annealing, estimating the edge-penalty parameter by MPL ("Maximum Pseudo-Likelihood") [2].

4.6 Experimental results

The first experiment presented here was carried out on the whole SPOT image. Classifications performed by the four TS-MRF solutions are shown in Figures 10-13, while the result of H-MAP is shown in Fig.9. We also present the confusion matrices, with the associated classwise accuracies (producer's and user's accuracies) in Tables 1-4, both for reference methods (found in [5]) and for TS-MRF algorithms. The global accuracy indicators for all these classifiers are gathered in Tab.5.

By inspection of the global indicators, we observe that contextual outperform non-contextual ones and, also, that TS-MRF methods in general perform better than reference, even considering only uncorrelated bands. However, global accuracy indicators hide deep differences of behavior between the classifiers, so we now proceed to a detailed class-based examination of results.

Water classification accuracy is very high for all the classifiers. For bare soils, all the results are also rather good in terms of producer's accuracy, which never drops below 87%, and definitely good for user's accuracy, always above 96%. Hence, these classes do not contaminate the others.

For other classes, results show significant variations. Urban areas accuracy ranges from 57% (MD) to 95% (TS.C) in terms of producer's accuracy and from 66% (ML) to 81% (DA) in terms of user's accuracy. With respect to this class, the better classifiers seem to be H-MAP and TS-MRF. In particular, in this case the use of correlated bands significantly improves the classification, as made clear by accuracy indicators or by the classification maps themselves. For example, a track which crosses the scene is well detected by TS.C and TS.C+ (see Figures 12 and 13), while is missed by all the other contextual classifications, including those not shown here (see again [5]). For the forest/heath class, contextual classifiers and ML provide user's and producer's accuracies that are slightly better than those of DA and MD.

Although all results are bad for meadows, due to a very high overlap between temporary and permanent meadows, the TS-MRF classifiers give much better results than the others. This remarkable case makes clear the benefits of using the proposed tree-structured model. As a matter of fact,

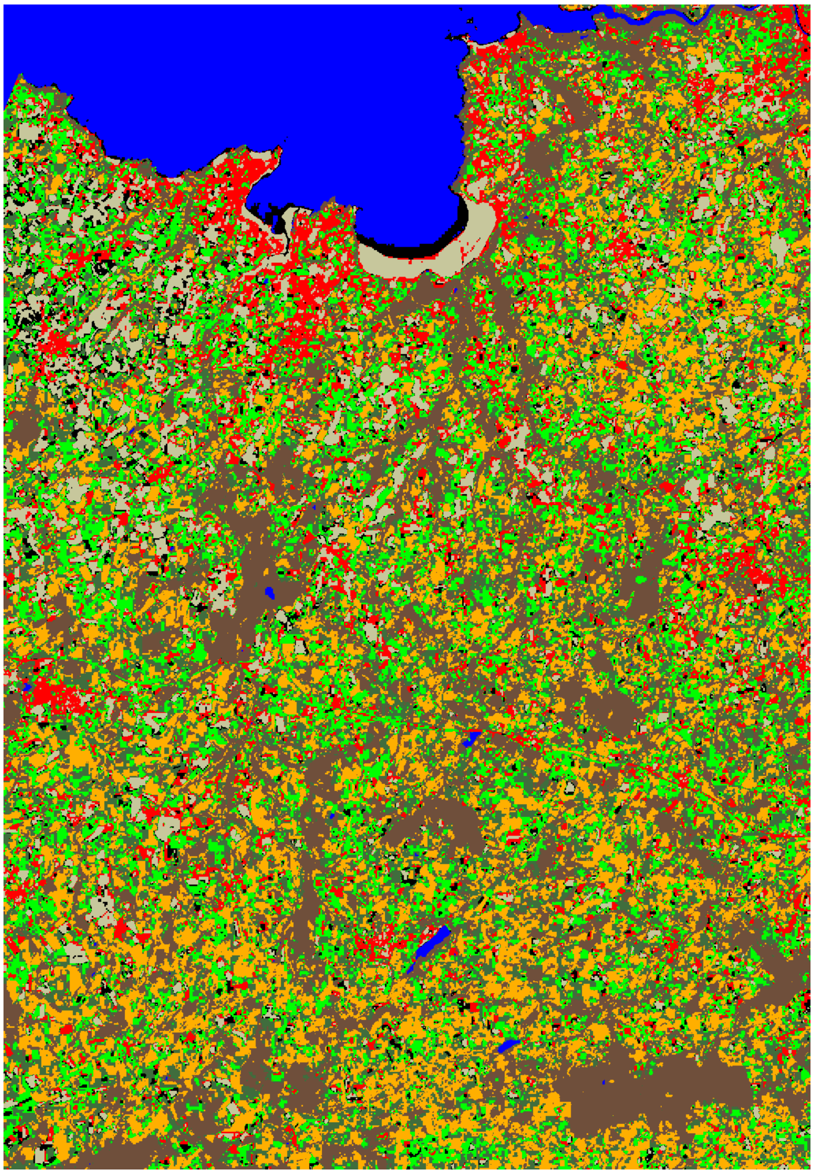


Figure 9: Classification by H-MAP

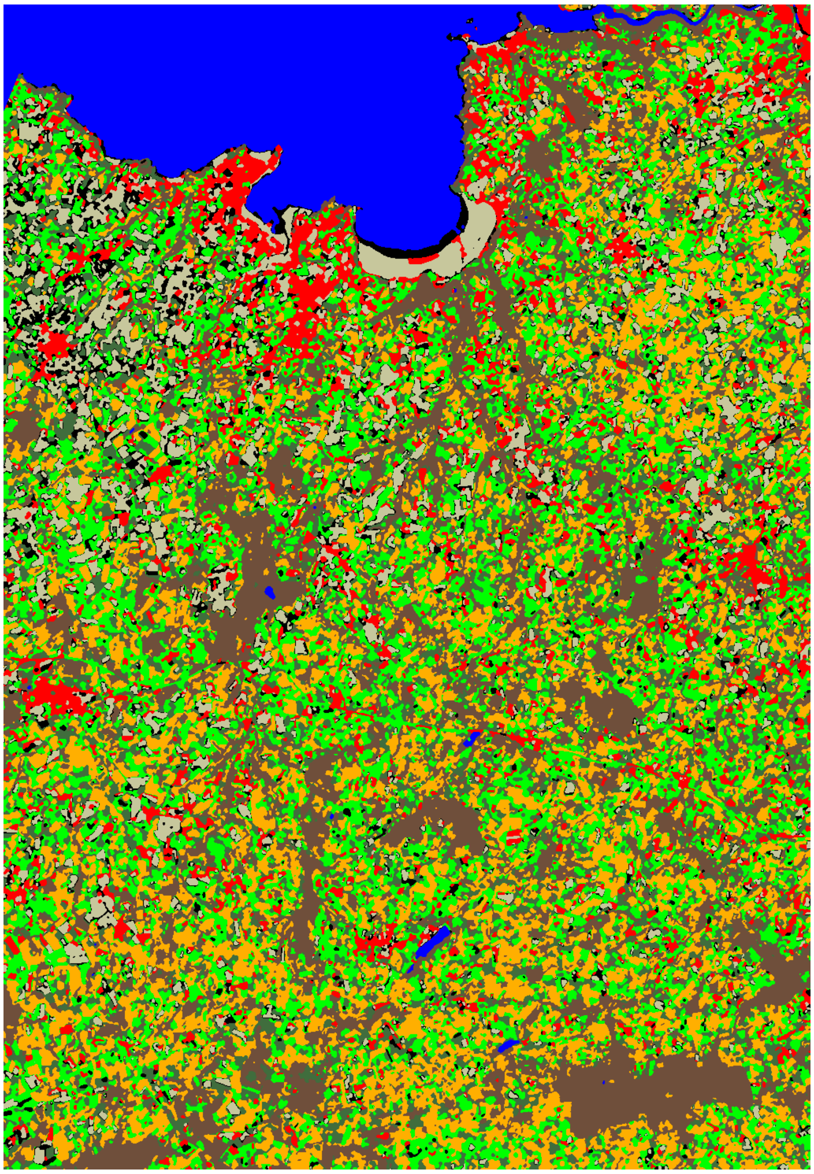


Figure 10: Classification by TS.U

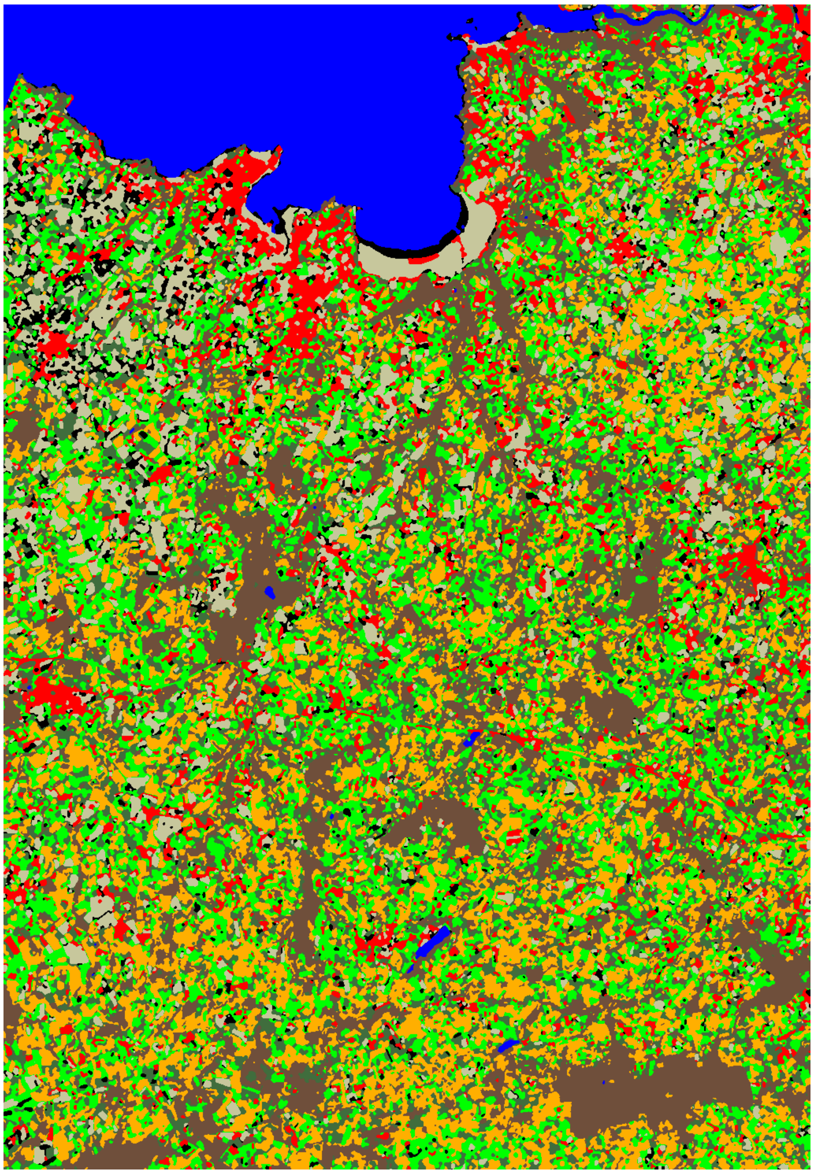


Figure 11: Classification by TS.U+

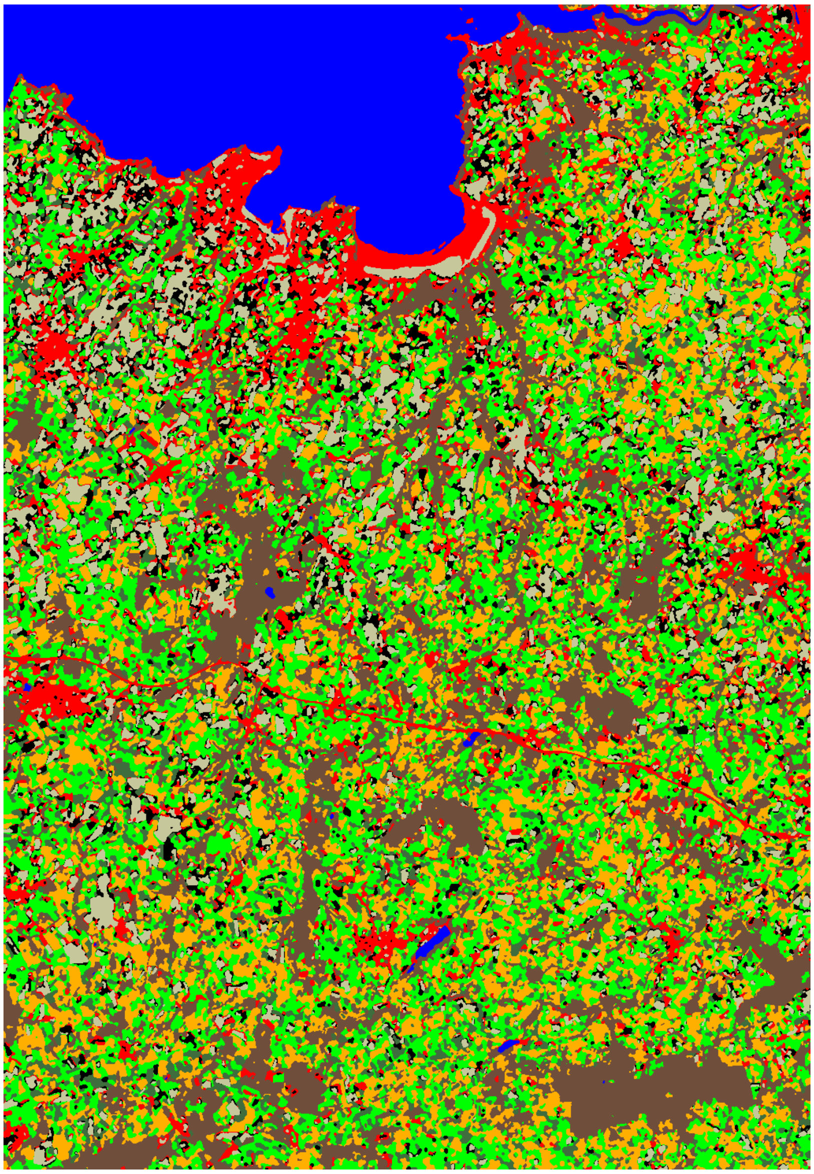


Figure 12: Classification by TS.C

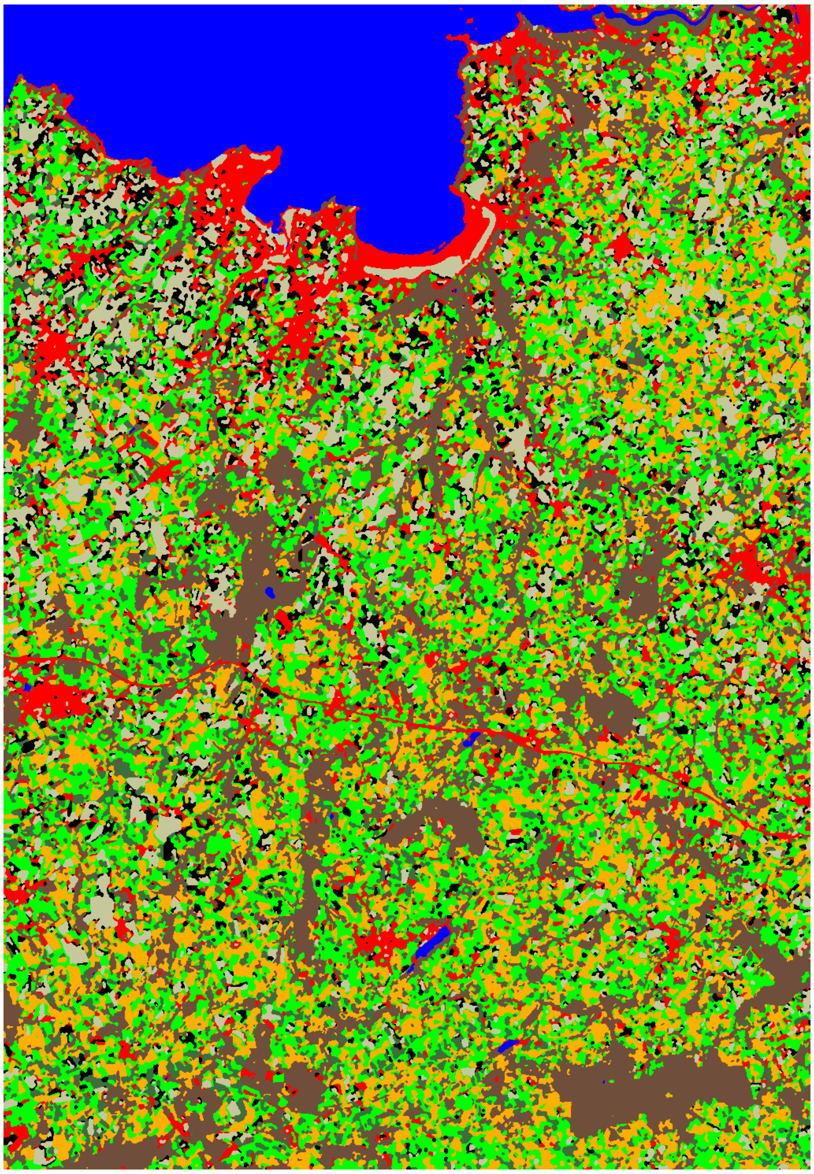


Figure 13: Classification by TS.C+

Discriminant Analysis

	water	bare soil	urban areas	forests/ heath	tempor. meadows	perman. meadows	veget- ables	corn	Total	User's accur.
water	**547**			12					559	97,9%
bare soils		**1327**	39						1366	97,1%
urban		43	**317**	1	12	11		5	389	81,5%
forests	8	8	28	**1473**	43			6	1566	94,1%
tem. mead.		7	44		**113**	98	3	10	275	41,1%
per. mead.				9	124	**47**	2	52	234	20,1%
vegetab.		105	40		27	18	**0**		190	0%
corn		1		96	71	220		**433**	821	52,7%
Total	555	1491	468	1591	390	394	5	506	5400	
Prod.'s acc.	98,6%	89,0%	67,7%	92,6%	29,0%	11,9%	0%	85,6%		

Minimum distance

	water	bare soil	urban areas	forests/ heath	tempor. meadows	perman. meadows	veget- ables	corn	Total	User's accur.
water	**545**		1	4					550	99,1%
bare soils		**1300**	46						1346	96,6%
urban		70	**268**	9	22	15		3	387	69,3%
forests	9		11	**1393**	26			3	1442	96,6%
tem. mead.	1	10	24	75	**114**	106	1	18	349	32,7%
per. mead.		1		55	75	**55**	3	54	243	22,6%
vegetab.		110	118		53	12	**1**	3	297	0,3%
corn				55	100	206		**425**	786	54,1%
Total	555	1491	468	1591	390	394	5	506	5400	
Prod.'s acc.	98,2%	87,2%	57,3%	87,6%	29,2%	13,9%	20,0%	84,0%		

Maximum Likelihood

	water	bare soil	urban areas	forests/ heath	tempor. meadows	perman. meadows	veget- ables	corn	Total	User's accur.
water	**528**			3					531	99,4%
bare soils		**1316**	49						1365	96,4%
urban		109	**283**		21	12		2	427	66,3%
forests	11			**1520**	2			26	1559	97,5%
tem. mead.		9	97	2	**138**	110	2	10	368	37,5%
per. mead.		3	9	7	119	**65**	3	36	242	26,9%
vegetab.	16	53	30	4	12	9	**0**		124	0%
corn		1		55	98	198		**432**	784	55,1%
Total	555	1491	468	1591	390	394	5	506	5400	
Prod.'s acc.	95,1%	88,3%	60,5%	95,5%	35,4%	16,5%	0%	85,4%		

Table 1: Confusion matrices for several spectral-based classifications performed on the whole image. The entry of ith row and jth column is the number of sample pixels from jth class that have been classified as belonging to the ith class.

Ising MRF

	water	bare soil	urban areas	forests/ heath	tempor. meadows	perman. meadows	veget- ables	corn	Total	User's accur.
water	**527**			2					529	99,6%
bare soils		**1368**	48						1416	96,6%
urban		88	**310**	1	19	17		4	439	70,6%
forests	12			**1534**				26	1572	97,6%
tem. mead.		4	96		**141**	109	5	5	360	39,2%
per. mead.		3	4	4	117	**61**		3	220	27,7%
vegetab.	12	27	10	5	11	3	**0**		72	0%
corn		1		45	102	204		**440**	792	55,6%
Total	555	1491	468	1591	390	394	5	506	5400	
Prod.'s acc.	94,9%	91,8%	66,2%	96,4%	36,2%	15,5%	0%	86,9%		

H-MAP

	water	bare soil	urban areas	forests/ heath	tempor. meadows	perman. meadows	veget- ables	corn	Total	User's accur.
water	**528**			4					532	99,3%
bare soils		**1341**	35						1376	97,5%
urban		91	**359**	1	14	20		2	487	73,7%
forests	11			**1525**	2			20	1558	97,9%
tem. mead.		8	68		**159**	113	5	17	370	43,0%
per. mead.		4	3	19	110	**54**		27	217	24,9%
vegetab.	16	46	3	3	15	1	**0**		84	0%
corn		1		39	90	206		**440**	776	56,7%
Total	555	1491	468	1591	390	394	5	506	5400	
Prod.'s acc.	95,1%	89,9%	76,7%	95,9%	40,8%	13,7%	0%	86,9%		

H-MPM

	water	bare soil	urban areas	forests/ heath	tempor. meadows	perman. meadows	veget- ables	corn	Total	User's accur.
water	**525**			9					534	98,3%
bare soils		**1353**	28					1	1382	97,9%
urban		89	**429**	1	12	20	2	2	555	77,3%
forests	14			**1549**				1	1564	99,0%
tem. mead.		20	6		**172**	130		21	349	49,3%
per. mead.			5	13	111	**31**	3	29	192	16,2%
vegetab.	16	27			4		**0**		47	0%
corn		2		19	91	213		**452**	777	58,2%
Total	555	1491	468	1591	390	394	5	506	5400	
Prod.'s acc.	94,6%	90,7%	91,7%	97,4%	44,1%	7,9%	0%	89,3%		

Table 2: Confusion matrices for several contextual-based classifications performed on the whole image.

TS.U

	water	bare soil	urban areas	forests/ heath	tempor. meadows	perman. meadows	veget- ables	corn	Total	User's accur.
water	**527**			1					528	99,8%
bare soils		**1343**	18					1	1362	98,6%
urban		94	**416**	1	7	17		2	537	77,5%
forests	13			**1518**				27	1558	97,4%
tem. mead.		6	17	1	**221**	117	5	10	377	58,6%
per. mead.				11	66	**63**		30	174	36,2%
vegetab.	15	4	17	11	19		**0**		106	0%
corn		44		48	77	197		**436**	758	57,5%
Total	555	1491	468	1591	390	394	5	506	5400	
Prod.'s acc.	94,9%	90,1%	88,9%	95,4%	56,7%	16,0%	0%	86,2%		

TS.U+

	water	bare soil	urban areas	forests/ heath	tempor. meadows	perman. meadows	veget- ables	corn	Total	User's accur.
water	**527**			1					528	99,8%
bare soils		**1358**	19					1	1378	98,5%
urban		94	**416**	1	8	17		2	538	77,3%
forests	13			**1518**				27	1558	97,4%
tem. mead.		6	17	1	**221**	117	5	10	377	58,6%
per. mead.		4		11	66	**63**		30	174	36,2%
vegetab.	15	29	16	11	18		**0**		89	0%
corn				48	77	197		**436**	758	57,5%
Total	555	1491	468	1591	390	394	5	506	5400	
Prod.'s acc.	94,9%	91,1%	88,9%	95,4%	56,7%	16,0%	0%	86,2%		

TS.C

	water	bare soil	urban areas	forests/ heath	tempor. meadows	perman. meadows	veget- ables	corn	Total	User's accur.
water	**543**			3					546	99,4%
bare soils		**1375**	10						1385	99,3%
urban		61	**443**	12	29	15		6	566	78,3%
forests	12			**1539**				26	1577	97,6%
tem. mead.		10		10	**259**	90	5	39	413	62,7%
per. mead.		2		2	7	**123**		69	203	60,6%
vegetab.		42	15		28	25	**0**	1	111	0%
corn		1		25	67	141		**365**	599	60,9%
Total	555	1491	468	1591	390	394	5	506	5400	
Prod.'s acc.	97,8%	92,2%	94,7%	96,7%	66,4%	31,2%	0%	72,1%		

Table 3: Confusion matrices for several classifications performed on the whole image by TS-MRFs.

TS.C+

	water	bare soil	urban areas	forests/ heath	tempor. meadows	perman. meadows	veget- ables	corn	Total	User's accur.
water	**543**			3					546	99,4%
bare soils		**1397**	13	1				3	1414	98,8%
urban		45	**444**	11	29	20		3	552	80,4%
forests	12			**1539**				26	1577	97,6%
tem. mead.		10		10	**259**	90	5	39	413	62,7%
per. mead.		2		2	7	**123**		69	203	60,6%
vegetab.		36	11		28	20	**0**	1	96	0%
corn		1		25	67	141		**365**	599	60,9%
Total	555	1491	468	1591	390	394	5	506	5400	
Prod.'s acc.	97,8%	93,7%	94,9%	96,7%	66,4%	31,2%	0%	72,1%		

Table 4: Confusion matrix for a TS-MRF classification performed on the whole image.

Methods	DA	MD	ML	ISING	H-MAP	H-MPM	TS.U	TS.U+	TS.C	TS.C+
τ	78,8%	75,9%	79,3%	81,1%	81,6%	83,5%	83,8%	84,1%	86,1%	86,5%
κ	73,8%	70,5%	74,3%	76,5%	77,1%	79,5%	79,9%	80,2%	82,7%	83,2%
τ^{norm}	55,7%	60,7%	55,2%	55,3%	56,3%	52,7%	57,8%	57,9%	60,7%	60,8%

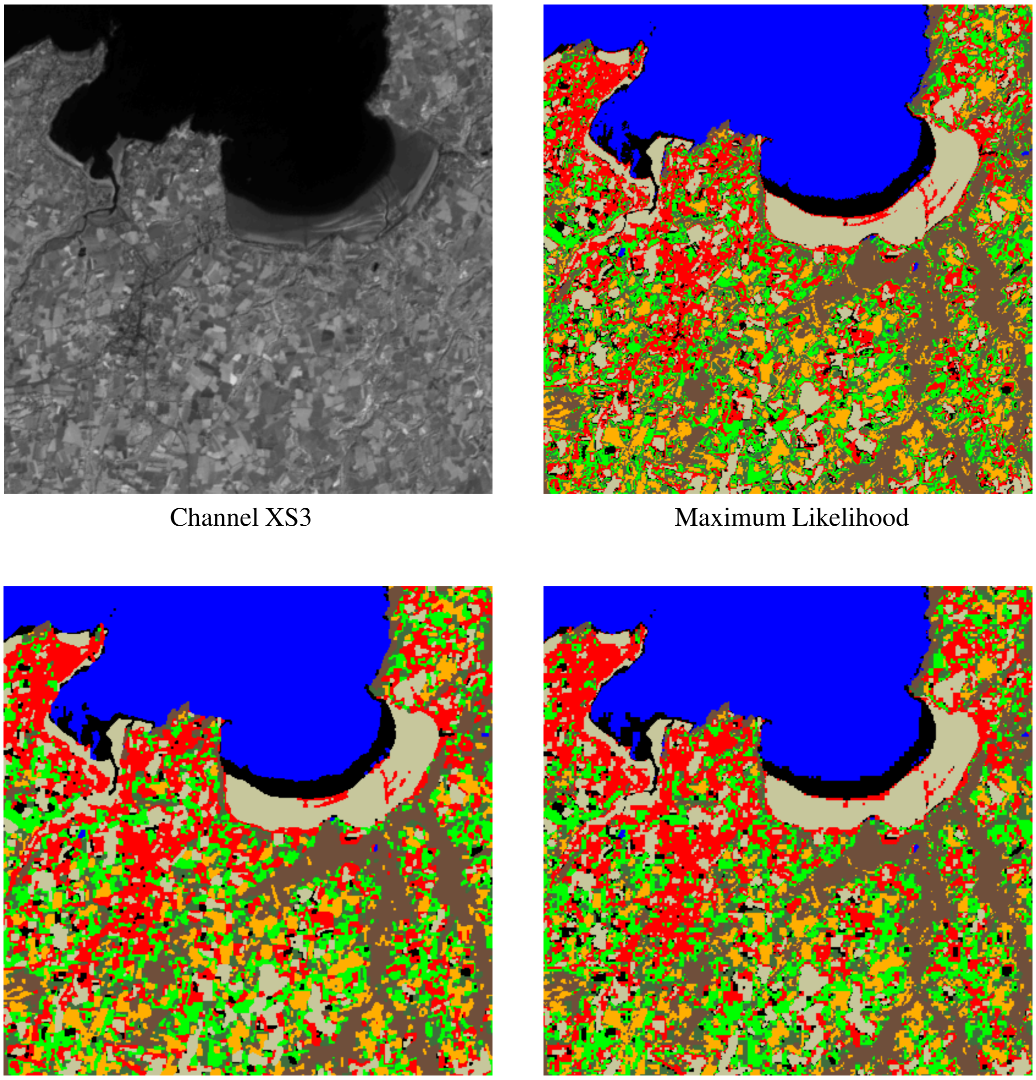
Table 5: Summary of the global accuracy indicators for the classifications performed on the whole image.

such two classes are associated in the tree with the same parent node (see Fig.8) and, as a consequence, are always kept together and finally split only once their shared region has been singled out by ancestor splits.

For vegetables, all results are very poor. Such a global failure should be mostly attributed to the insufficient amount of pixels available for this class in the learning and validation sets. Corn regions are rather well extracted by all methods with a minimum producer's accuracy of 72% but are quite contaminating, with a user's accuracy of the order of 55%.

The second experiment presented here concerns the small portion (400×400) of the whole scene (see Fig.14). We performed this experiment in order to carry out a comparison also with the variational methods proposed in [32] which were tested on this image. In Figures 14 and 15, classifications carried out by TS-MRF methods and reference classifiers are shown, while global accuracy indicators are gathered in Table 6. For the two variational methods M1X and M2X, only the overall accuracy τ is available, and based on such indicator the TS-MRF classifiers work better than both. As an aside, the comparative analysis made above with respect to H-MAP, ISING and ML, is confirmed on this small image.

By inspection of the indicators associated to both the experiments proposed it seems that the use of an adaptive Ising as basic component does not improve the algorithm. However, we point out that one of the three classes interested by the adaptive model is the vegetable class, whose accuracy analysis is completely unreliable. In addition, we expect to observe significant differences between TS and TS+ in the classification of mixed pixels (see [15]), that is boundary cells where a class change occurs. But the test set (Fig.7) is composed of a collection of homogeneous, "unconnected",



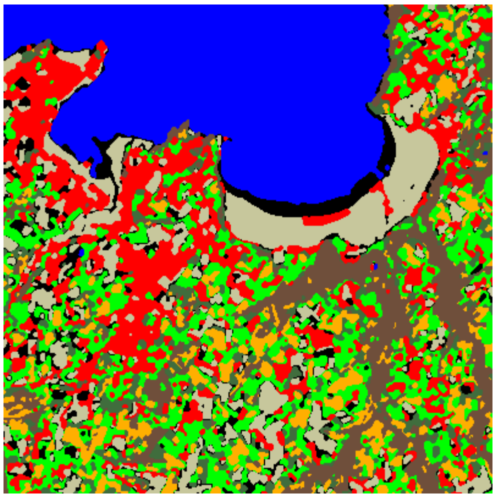
Channel XS3 | Maximum Likelihood

Ising MRF | H-MAP

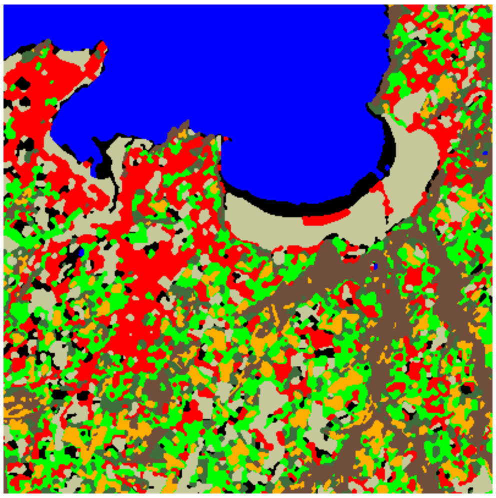
Figure 14: Small SPOT image (400×400) and some classifications.

Methods	ML	ISING	H-MAP	M1X	M2X	TS.U	TS.U+	TS.C	TS.C+
τ	65,1%	66,0%	70,5%	71,3%	70,0%	73,3%	73,3%	75,7%	75,8%
κ	57,9%	58,8%	64,2%	–	–	67,4%	67,5%	70,2%	70,2%
τ^{norm}	51,5%	45,3%	50,1%	–	–	52,8%	52,8%	55,8%	54,5%

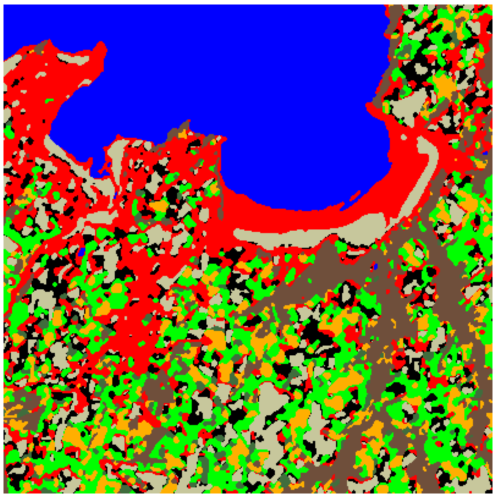
Table 6: Summary of the global accuracy indicators for the classifications performed on the small image.



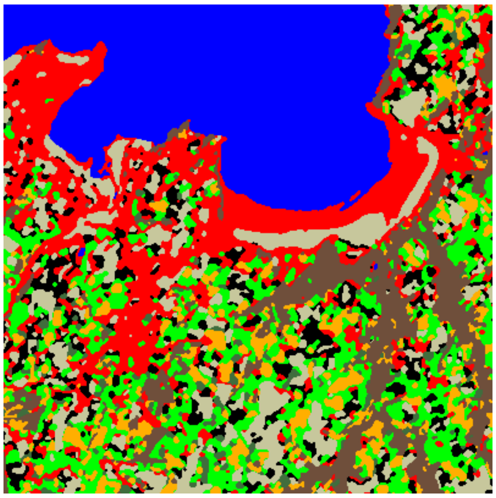
TS.U



TS.U+



TS.C



TS.C+

Figure 15: Classifications of the small SPOT image by TS-MRF algorithms.

rectangular regions, with no mixed cells, so that no appreciable improvements can be observed. Nevertheless, by visual inspection of classifications (for instance in Fig.15) one can see how the boundary fragmentation is less frequent for TS+ with respect to TS; the most perceptible change is for vegetables (shown as black).

Finally, we observe, for both the experiments, a noticeable difference between the algorithms which use the inter-band correlation, TS.C and TS.C+, and all the others. In fact, a large region along the sea front is classified as urban area by TS.C(+), while all other classifiers put it in the bare soils class, as would seem reasonable. This anomalous behaviour is certainly due to the likelihood model, significantly changed by the use of correlated channels, which indicates the urban class as more likely. To prove this, we segmented the image by a ML classifier with correlated channels and, eventually, observed the same anomaly. Of course, in this sense, the choice of the learning set is crucial. Likewise, in same area, a ground strip corresponding to vegetables for uncorrelated models is classified as water by both TS.C and the correlated-bands ML classifier, further validating the considerations made above.

5 Unsupervised image segmentation

5.1 The unsupervised TS-MRF algorithm

The nature of the TS-MRF model suggests to address the unsupervised segmentation problem by using the recursive optimization procedure with the addition of a stopping condition at each newly created leaf in order to decide on the growth of the tree. As a result, the number of classes in the image (*cluster validation problem*), that is the number of terminal nodes of the tree, is automatically detected, as a byproduct of the tree structure itself.

We now describe the algorithm by first giving the initialization, focusing on the root (node 1), and then the generic step with reference to a given tree.

At the beginning, we consider the following two hypotheses[5]:

$$\begin{cases} H_0: & T = \{1\}, & X^T = \emptyset \\ H_1: & T' = \{1,2,3\}, & X^{T'} = \widehat{x}^1 \end{cases} \tag{41}$$

The first hypothesis corresponds to the case in which the whole image, associated with the root node ($\mathcal{S}^1 = \mathcal{S}, y^1 = y$), is represented as a single region. Therefore, the observed data are described by a single distribution $p(y^1)$, whose model is known except for some parameters θ^1 that must be estimated from the data themselves. Of course, in this case the TS-MRF is empty, and all sites have the same label attached. This is the only possible configuration and in this sense we define $p(x^{\{1\}}) = 1$, and also write the data distribution as $p(y^1|\theta^1)$ to make explicit that y^1 is described through the single set of parameters θ^1 attached with node 1.

The second hypothesis corresponds to the case in which the image is represented by two regions. To single out such regions, a binary MRF X^1 is defined on $\mathcal{S}^1$, with potentials $V_c^1(\cdot)$ that are completely specified except for some parameters β_1. The MAP estimate of the MRF $\widehat{x}^1$, with probability $p(\widehat{x}^1)$, divides the image into two new regions, $\mathcal{S}^2 = \{s \in \mathcal{S}^1 : \widehat{x}^1_s = 2\}$ and $\mathcal{S}^3 = \{s \in \mathcal{S}^1 : \widehat{x}^1_s = 3\}$, with their associated data y^2 and y^3.

Recall that we assume, as it is usually done, that the data are conditionally independent given the labels, and hence their description factors out as $p(y^1|x^1) = p(y^2|\theta^2)p(y^3|\theta^3)$.

At this point, we compare the two statistical descriptions of the image, based on a single-class model (tree T) or a two-class model (tree T'), by checking the condition

$$G^1 = \frac{p(\widehat{x}^{T'})p(y|\widehat{x}^{T'})}{p(\widehat{x}^{T})p(y|\widehat{x}^{T})} > 1 \tag{42}$$

which, specialized for $T = \{1\}$, becomes

$$G^1 = \frac{p(\widehat{x}^1)}{1} \times \frac{p(y^1|\widehat{x}^1)}{p(y^1|\theta^1)} > 1 \tag{43}$$

If the test succeeds, namely the split gain G^1 is greater than 1, the two-region description fits the data better and the procedure goes on, otherwise it stops and the single-region description is accepted.

[5]Now, since the tree structure is not given, we make explicit reference to it in notation, for instance using X^T to refer at the overall field X, where T is a tree structure, to distinguish from X^t that is a binary field associated to node t.

Let us now consider a generic tree T, that has been temporarily accepted as our structure, with associated TS-MRF X^T, and let τ be a leaf of T that we are testing for a possible split. The two hypotheses under test are then

$$\begin{cases} H_0: & T, & X^T = \widehat{x}^T \\ H_1: & T' = \text{split}(T,\tau), & X^{T'} = (\widehat{x}^T, \widehat{x}^\tau) \end{cases} \tag{44}$$

where tree $T' = \text{split}(T,\tau)$ is identical to T except for node τ which generates two new leaves, 2τ and 2τ+1, becoming an internal node itself. To explicitly test (42) for the general case, remember that $p(x^T) = \prod_{t\in\overline{T}} p(x^t|x^{\omega(t)})$, and that $p(x^t|x^{\omega(t)}) = p(x^t|\mathcal{S}^t)$. Moreover $p(y|x^T) = \prod_{t\in\widetilde{T}} p(y^t|\theta^t)$. Therefore, we can write

$$\begin{aligned} p(\widehat{x}^T) &= \prod_{t\in\overline{T}} p(\widehat{x}^t|\mathcal{S}^t) \\ p(\widehat{x}^{T'}) &= \prod_{t\in\overline{T}'} p(\widehat{x}^t|\mathcal{S}^t) = p(\widehat{x}^\tau|\mathcal{S}^\tau) \prod_{t\in\overline{T}} p(\widehat{x}^t|\mathcal{S}^t) \\ p(y|\widehat{x}^T) &= \prod_{t\in\widetilde{T}} p(y^t|\theta^t) = p(y^\tau|\theta^\tau) \prod_{t\in\widetilde{T},t\neq\tau} p(y^t|\theta^t) \\ p(y|\widehat{x}^{T'}) &= \prod_{t\in\widetilde{T}'} p(y^t|\theta^t) = p(y^\tau|\widehat{x}^\tau) \prod_{t\in\widetilde{T},t\neq\tau} p(y^t|\theta^t) \end{aligned} \tag{45}$$

and the split gain simplifies to

$$\begin{aligned} G^\tau &= \frac{p(\widehat{x}^\tau|\mathcal{S}^\tau)}{1} \times \frac{p(y^\tau|\widehat{x}^\tau)}{p(y^\tau|\theta^\tau)} \\ &= p(\widehat{x}^\tau|\mathcal{S}^\tau) \times \frac{p(y^{2\tau}|\theta^{2\tau})p(y^{2\tau+1}|\theta^{2\tau+1})}{p(y^\tau|\theta^\tau)} \end{aligned} \tag{46}$$

It should be noted that the test depends exclusively on region $\mathcal{S}^\tau$. In fact, given $\widehat{x}^T$ the maximization process operates only on x^τ, and the MAP problem reduces to

$$\widehat{x}^\tau = \arg\max_{x^\tau} p(x^\tau)p(y^\tau|x^\tau) \tag{47}$$

which is completely local to node τ. If the test succeeds, the growth of the tree and of the associated segmentation continues in a similar way for each newly created leaf, as if each one were the root of a new tree. Therefore, the tree growing process is accurately described by a recursive procedure, which can go on in parallel for each node.

The likelihood part in Eq.47, $p(y^\tau|x^\tau)$, is not expressed by Eq.40 like in the supervised case, because the number of descendant leaves of τ is now unknown. Therefore, we suppose to have only one descendant for each of the two children of τ, thus achieving a greedy optimization procedure.

The ratio G^τ, referred to as the split gain, accounts for the gain in description efficiency arising from the split of leaf τ. This interpretation becomes more compelling if we take the logarithm of

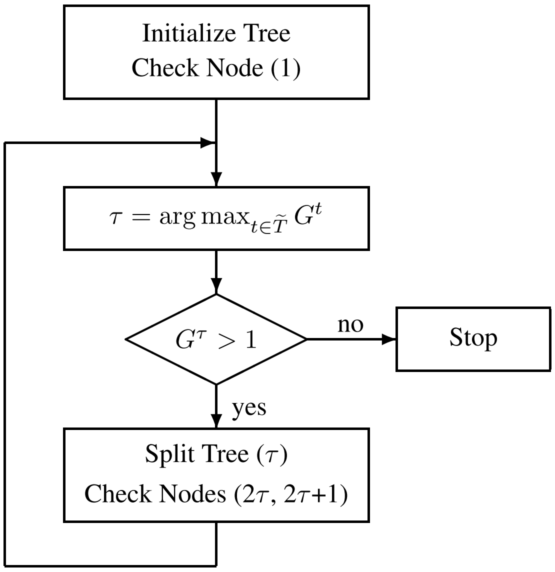


Figure 16: High-level flow chart of the segmentation algorithm.

G^τ and regard it as the difference $\log G^\tau = I(T) - I(T')$ between the self-information associated with each of the competing TS-MRF's[6]. If the self-information is a good indicator of the description complexity, then a positive log split gain indicates that the new description of the observed data is "simpler" than the preceding one, and hence preferable (according to Occam's razor). In more detail, a split has always a cost, $p(\widehat{x}^t) < 1$, due to the need of describing the segmentation $\widehat{x}^t$, but also a value, $p(y^t|\widehat{x}^t)/p(y^t|\theta^t) > 1$, because the data are more accurately represented, in each new segment, by their local parameters. A positive $\log G^t$ indicates that the overall benefits outweight the cost.

Fig.16 shows a high-level flow chart of the segmentation algorithm. To improve readability, the procedure is sequential rather than recursive, and only one leaf at a time is split, the one with the largest split gain (the experiments will follow this convention as well).

- In the initialization step, the tree is defined as consisting of the sole root ($T = \{1\}$); the whole image is associated with it ($\mathcal{S}^1 = \mathcal{S}, y^1 = y$), and the vector of parameters $\widehat{\theta}^1$ is estimated; of course, the TS-MRF is empty ($X^T = \emptyset$).

- In the procedure CheckNode(t), the binary MRF X^t is defined on $\mathcal{S}^t$, the MAP realization $\widehat{x}^t$ is estimated together with its parameters $\widehat{\beta}^t$, and the split gain G^t is evaluated. If $G^t > 1$ this node will be split sooner or later.

[6]This discussion is only to gain insight about the meaning of the split gain, and there is no attempt to be rigorous.

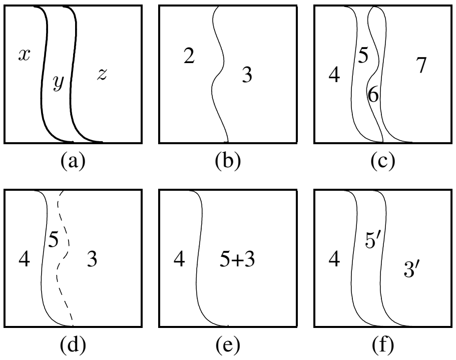


Figure 17: Using a split-only approach, the synthetic image (a) can be erroneously segmented in 4 classes (c). Allowing for region merging, the correct solution is obtained (f).

- SplitTree(t) updates the structure of the tree by moving t from $\widetilde{T}$ to $\overline{T}$, and generating two new leaves $2t$ and $2t$+1; to each one of such new nodes the proper quantities ($\mathcal{S}^{2t}, y^{2t}, \widehat{\theta}^{2t}$, etc.) are associated (they were evaluated during the CheckNode step).

5.2 A split-and-merge solution

As was remarked in the previous section, in the unsupervised case a split is carried out by modelling the data of the two regions to extract as corresponding to terminal nodes, while in the supervised case they are supposed to correspond to two sub-trees (with known structure). This fact leads to a greedy algorithm in the first case where a single object can be over-segmented into two or more regions, with introduction of false contours.

The example of Fig.17 illustrates this problem: in (a) we have a synthetic image with three distinct regions; because only binary segmentations are possible, it can happen (given the data model) that the central region y is split in half among the two newly formed segments, 2 and 3 (b); after two more splits, we have the situation shown in (c), where two different segments, 5 and 6, correspond to two adjacent parts of the same region y, a clear failure of the algorithm.

Various solutions can be envisaged for this problem, such as, for example, allowing a variable number of children for each node. A simple and fast solution which keeps the binary structure, to allow for the reshaping of region boundaries by means of a split-and-merge procedure was presented in [14]. To gain insight about the rationale of this approach, consider again Fig.17: after the split of node 2, we have leaves (and regions) 4, 5, and 3, as shown in (d); at this point, we can try to merge each of the new nodes (4 and 5) with all existing leaves (only node 3 in this case) and assess the possible gain: the merging of nodes 5 and 3 should indeed provide a gain, because region y will then belong to a single class. In this case, the merging takes place (e); this merged region (5+3) will eventually be split again resulting in regions 5' and 3', and hence the correct segmentation shown in part (f). The split-and-merge alternating procedure is driven making use of two indicators: the split

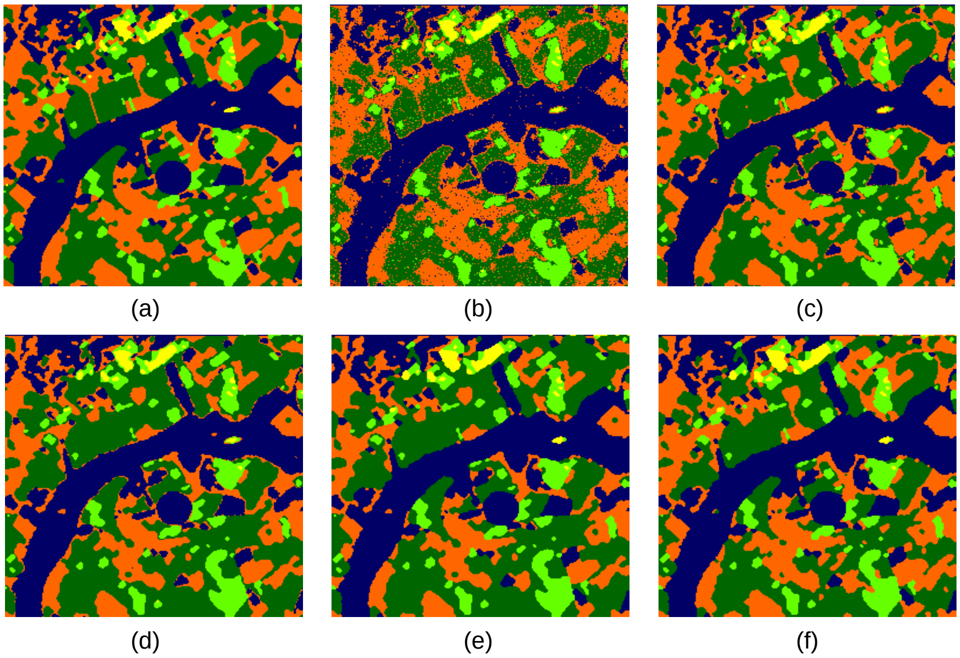


Figure 18: Synthetic multispectral image: ground truth (a), and segmentation maps produced by MD (b), Ising MRF (c), TS-MRF (d), adaptive TS-MRF (e), adaptive TS-MRF with merging (f).

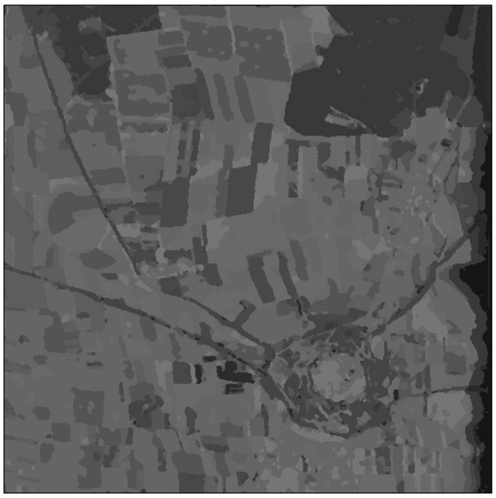
gain defined above, which controls the splitting, and a dual parameter, referred to as the merge gain (see [14] for more details), which drives the merging.

In Fig.18 a comparison between several TS-MRF versions, but also with standard techniques, is shown. In Fig.18, together with the ground-truth, we have the segmentation maps produced by various algorithms, that is, minimum distance (b), Ising-MRF (c), TS-MRF (d), TS-MRF with adaptive Ising components (e), and TS-MRF with adaptive Ising component and merging (f). All algorithms correctly stop at $K = 5$ (this was an input parameter for MD), which is reasonable for such a simple image.

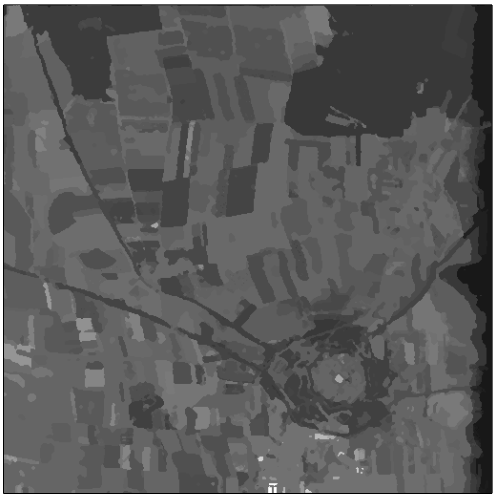
5.3 Unconnected regions splitting

Describing a field of data by means of two sets of parameters, rather than one, is always convenient in terms of accuracy. Therefore, the only reason for not splitting a region in two is the increased description complexity, related to the number of edges between the pixels of the two new subregions.

(a)



(b)



(c)

Figure 19: A comparison of TS-MRF algorithms with (c) or without (b) splitting of unconnected regions on a GER image (a).

In the splitting procedure used in unsupervised TS-MRF algorithm, quite often children regions are composed of several spatially distinct areas[7]. Splitting such areas from one another, no matter how similar they are, is necessarily associated with a positive gain since no new edges (and edge penalties) are introduced in the process. Actually, assuming all parameters, including the split gain, are correctly estimated, such kind of splits should occur anyway, sooner or later, while the segmentation tree grows.

Based on this consideration a variation of the TS-MRF algorithm has been proposed recently [16], which enforces, after each binary split, the separation of non-connected areas to form new nodes. The separation of non-connected regions presents several advantages. First of all, it is self evident that segmentation becomes very fast, the bulk of processing being concentrated in the first split, for which multi-resolution techniques [3, 8, 20, 33] can also be used. In addition, by removing the influence of data from distant areas in the parameter estimation phase, the adaptation to local statistics becomes more accurate. A further advantage lies in the opportunity to associate meaningful geometrical and topological features with each node/region. Experiments also show (and we are still working to explain this behavior) that the modified algorithm succeeds in extracting small-size details that do not appear when only binary splits are used.

On the down side, in this approach focused on segments rather than classes, some useful information is lost regarding the meaning of each region. In other words, non-connected regions characterized by the same land cover do not share anymore a common label, and present different sets of parameters, possibly very similar to one another, that must be recorded separately. Of course, such relationships can be recovered with some more efforts in a further phase of processing.

Fig.19 shows the result of a segmentation experiment. A 512x512 section of an 8-band GER remote sensing image, one band of which is shown in part (a), is segmented by both the original TS-MRF algorithm and the last version with separation of non-connected components. The resulting segmentation maps, projected on the reference bands, are shown in parts (b) and (c). It is clear that, with the second algorithm, many more small-size regions are captured (some brighter ones are especially visible in the bottom part of the image) and many false contours are avoided, with no need of expensive split-and merge procedures.

5.4 Computational complexity

We show here a comparison, in terms of computational burden, between the standard TS-MRF and the reference algorithm, which is a K-class Ising MRF model. To this end, we consider again the GER image described above (a 512×512 pixel section), which is segmented using both the algorithms which are referred to simply as “TS” and “flat” algorithms.

Fig.20 reports CPU time as a function of the number of classes when a single-band image is considered. Contrary to our intuition, in the supervised case, when the number of classes is known *a priori*, the flat algorithm is almost always faster than the TS one. As a matter of fact, the tree structure entails the additional burden of dealing with irregular lattices, which calls for more complex data structures and additional controls. On the other hand, the simple MRF model used in both cases, and the suboptimal ICM algorithm which guarantees a quick convergence, reduce the potential

[7] Indeed, this also happens with K-ary segmentations.

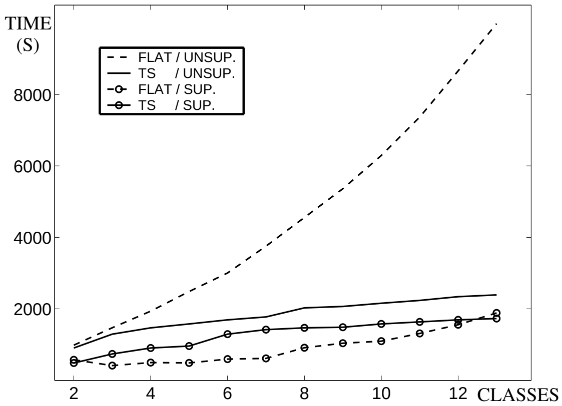


Figure 20: CPU time as a function of the number of classes, for “flat” and “TS” algorithms. In the unsupervised case, the TS algorithm is clearly preferable.

computation gain of using only binary fields. Such an advantage could become significant with more sophisticated MRF models and optimization algorithms.

However, if we consider the unsupervised case, the picture changes completely. With the flat algorithm, to obtain and accept a K-class map, all independent segmentations with $k = 2, \ldots, K+1$ classes are needed in order to evaluate their validity measure. With the TS algorithm, instead, all segmentations are nested, and the tests required to stop the growth are all local: therefore, given the K-class map, only another layer of splits is required, one for each leaf of the tree, with a nearly constant additional complexity with respect to the supervised case. As a consequence, using the TS-MRF model becomes more and more convenient as the number of classes grow (which is often the case for large images).

This advantage increases when we consider multiple bands. Fig.21 reports the CPU time as a function of the number of bands in the image for both the flat and TS algorithms when 8 classes are considered (supervised case). Complexity grows quite fast for the flat algorithm, and it soon exceeds that of the TS algorithm, which in turn remains almost constant. The effects of multiple bands on the complexity of the *unsupervised* case can be easily extrapolated without further experiments.

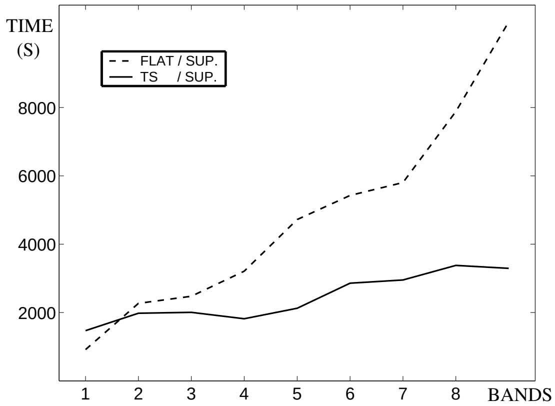


Figure 21: CPU time as a function of the number of bands, for “flat” and “TS” algorithms. The TS algorithm is more and more convenient as the number of bands increases.

6 Conclusions

In this work, we have presented a tree-structured MRF prior model useful for remote-sensing image segmentation and classification. Such a model presents several interesting properties.

First, it allows the generalization of each MRF model whose clique potentials depend on the configuration of the edges in the clique but are “blind” with respect to the color-transition of these edges, to a correspondent non-blind model. This new model (a TS-MRF), which includes also hierarchical constraints between classes, requires a number of parameters equal to $M \times (K-1)$ where M is the number of parameters of the reference blind model and K is the number of classes.

The imposed structural constraints between classes lead to a recursive factorization of the joint distribution in $K-1$ terms, each corresponding to an internal node of the tree and involving a corresponding prior parameter sub-set. As a consequence, a recursive procedure can be used to perform estimation-maximization avoiding further computational burden. Indeed, the algorithm proves experimentally to be faster then the basic reference model, especially when the number of bands becomes large, because all the recursive steps are simply “binary” segmentations that operate on smaller and smaller regions as the tree grows. Therefore, by using a larger parameter set in the context of a structured model, we increased the description capability of the model keeping unaltered, or even reducing, the computational complexity.

The output of the algorithm is not just the segmentation, but the whole tree, with its temporal development and all parameters associated with the individual nodes. Hence, given a K-leave segmentation tree, all intermediate trees and associated segmentation maps can be obtained by pruning back nodes in reverse order [30].

The availability of SPOT image with ground-truth, provided by the Costel laboratory (University of Rennes 2), allowed us to carry out experiments in supervised segmentation and to compare results with various other MRF-based and variational classifiers. The results show that, on this example, the TS-MRF classifiers outperform all reference techniques considered, both in terms of global accuracy and, even more, in terms of class-wise accuracy.

Because of its recursive structure, the TS-MRF model also addresses the cluster validation problem of unsupervised segmentation, through the use of a stopping condition local to each node. In this case, however, a *greedy* likelihood model is used for each node leading to a loss of accuracy. In order to improve the unsupervised TS-MRF algorithm, we introduced node merging and the splitting of unconnected regions, with some beneficial effects. Nevertheless, this is still an open problem to be addressed in future work.

Acknowledgments

The authors would like to thanks the GSTB ("Groupement Scientifique pour le développement de la Télédétection en Bretagne"), the COSTEL ("Climat, Occupation des Sols par TELédétection") laboratory, University of Rennes II, the VISTA ("VIsion Spatio-Temporelle et Active") project of IRISA Rennes for the ground-truth provided and for several comparative classification results.

References

[1] H.Akaike, "A new look at the statistical model identification," *IEEE Trans. Automatic Control*, vol.19, pp.716-722, Dec. 1974.

[2] J.Besag, "On the statistical analysis of dirty pictures," *Journal of the Royal Statistical Society*, Series B 48, pp.259-302, 1986.

[3] C.A.Bouman, M.Shapiro, "A multiscale random field model for Bayesian image segmentation," *IEEE Transactions on Image Processing*, pp.162-177, Mar. 1994.

[4] L.Breiman, J.H.Friedman, R.A.Olshen, C.J.Stone, *Classification and regression trees*, Wadsworth, Belmont (CA), 1984.

[5] A.Chardin, *Modèles énergétiques hiérarchiques pour la résolution des problèmes inverses en analyse d'images – Application à la télédétection*, PhD thesis, University of Rennes, France, Jan. 2000.

[6] A.Chardin, P.Pérez, "Semi-iterative Inferences with Hierarchical Energy-based Models for image analysis," EMMCVPR'99, Springer, Lecture Notes in Computer Science, Volume 1654, pages 83-98, York, England, July 1999.

[7] A. Chardin, P. Pérez, "Unsupervised Image Classification with a Hierarchical EM Algorithm," *Proc. Int. Conf. on Computer Vision*, ICCV'99, Vol.2, pp.969-974, Kerkyra, Greece, September 1999.

[8] H.Choi, R.G.Baraniuk, "Multiscale image segmentation using Wavelet-Domain Markov models," *IEEE Transactions on Image Processing*, vol.10, No.9, pp.1309-1321, Sep. 2001.

[9] R.G.Congalton, "A Review of assessing the accuracy of classifications of remotely sensed data," *Remote Sensing of Environment*, 37(1), pp. 35-36, 1991.

[10] A.P.Dempster, N.M.Larid, D.B.Rubin, "Mixtures densities, maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Stat. Society*, 39(1):1-38, 1977.

[11] X.Descombes, J.-F. Mangin, E.Pechersky, M.Sigelle, "Fine structures preserving Markov model for image processing," in *Proc. 9th SCIA '95*, Uppsala, Sweden, pp.349-356.

[12] X.Descombes, R.D.Morris, J.Zerubia, M.Berthod, "Estimation of Markov random field prior parameters using Markov chain Monte Carlo maximum likelihood," *IEEE Transaction on Image Processing*, vol.8, No.7, pp.954-963, July 1999.

[13] R.O.Duda, D.G.Stork, P.E.Hart, *Pattern Classification and Scene Analysis*, Wiley, 2000.

[14] C.D'Elia, G.Poggi, G.Scarpa, "Advances in the segmentation and compression of multispectral images," *Proc. 2001 IEEE International Geoscience and Remote Sensing Symposium*, vol.6, pp.2671-73, Sidney (AUS), July 2001.

[15] C.D'Elia, G.Poggi, G.Scarpa, "An adaptive MRF model for boundary-preserving segmentation of multispectral images," *Proc. XI European Signal Processing Conference*, Toulouse (F), Sept. 2002.

[16] C.D'Elia, G.Poggi, G.Scarpa, "Sequential Bayesian segmentation of remote sensing images," *Proc. IEEE International Conference on Image Processing*, Barcelona (Spain), Sept. 2003.

[17] C.D'Elia, G.Poggi, G.Scarpa, "A Tree-Structured Markov random field model for Bayesian image segmentation," *IEEE Transactions on Image Processing*, vol.12, no.10, pp.1259-1273, October 2003.

[18] S.Geman, D.Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp.721-741, Nov.1984.

[19] L.Hubert-Moy, A.Cotonnec, L.Le Du, A.Chardin, P.Pérez, "A comparison of parametric classification procedures of remotely sensed data applied on different landscape units," *Remote Sensing of Environment*, 75(2), pp.174-187, February 2001.

[20] Z.Kato, "Multiresolution Markovian modeling for computer vision. Application to SPOT image segmentation," PhD thesis, University of Nice-Sophia Antipolis, France, 1994.

[21] Z.Kato, M.Berthod, J.Zerubia, "A hierarchical Markov random field model and multitemperature annealing for parallel image classification," Graphical Models and Image Processing, vol.58, pp.18-37, Jan. 1996.

[22] S.Lakshmanan, H.Derin, "Simultaneous parameter estimation and segmentation of Gibbs random field using simulated annealing," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol.11, no.8, August 1989.

[23] J.M.Laferté, P.Pérez, F.Heitz, "Discrete Markov image modeling and inference on the quadtree," *IEEE Transactions on Image Processing*, pp.390-404, Mar.2000.

[24] D.Landgrebe, "Hyperspectral image data analysis," *IEEE Signal Processing Magazine*, vol.19, pp.17-28, Jan 2002.

[25] S.Z.Li, *Markov random field modeling in image analysis*, Springer-Verlag, 2001.

[26] M.Mignotte, C.Collet, P.Pérez, P.Bouthemy, "Sonar image segmentation using an unsupervised hierarchical MRF model," *IEEE Transactions on Image Processing*, pp.1216-1231, July 2000.

[27] A.Mohammad-Djafari, "Joint estimation of parameters and hyperparameters in a Bayesian approach of solving inverse problems," *Proc. IEEE International Conference on Image Processing*, vol.2, pp.473-476, 1996.

[28] G.Poggi, A.R.P.Ragozini, "Image segmentation by tree-structured Markov random field," *IEEE Signal Processing Letters*, pp.155-157, July 1999.

[29] A.Richards, *Remote sensing digital analysis: an introduction*, 2nd ed. Springer Verlag, Berlin Heidelberg, 1993.

[30] P.Salembier, L.Garrido, “Binary partition tree as an efficient representation for image processing, segmentation, and information retrieval,” *IEEE Transactions on Image Processing*, pp.561-576, April 2000.

[31] F.Salzenstein, W.Pieczynski, “Global and local methods of unsupervised Bayesian segmentation of images,” Graphical Models and Image Processing, vol.59, pp.205-220, 1997.

[32] C.Samson, L.Blanc-Féraud, G.Aubert, J.Zerubia, “Deux modèles variationnels pour la classification d’images multispectrales,” *Traitement du Signal 2001*, vol.18, no.5-6, pp.345-367.

[33] A.S.Willsky, “Multiresolution Markov models for signal and image processing,” *Proceedings of the IEEE*, vol.90, No.8, pp.1396-1458, Aug. 2002.

[34] G.Winkler, *Image analysis, random fields and dynamic Monte Carlo methods*, Springer-Verlag, 1995.

[35] J.Zhang, J.W.Modestino, D.A.Langan, “Maximum-likelihood parameter estimation for unsupervised stochastic model-based image segmentation,” *IEEE Trans. on Image Processing*, vol.3, no.4, pp.404-420, July 1994.

Unité de recherche INRIA Sophia Antipolis
2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex (France)

Unité de recherche INRIA Lorraine : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex (France)
Unité de recherche INRIA Rennes : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex (France)
Unité de recherche INRIA Rhône-Alpes : 655, avenue de l'Europe - 38330 Montbonnot-St-Martin (France)
Unité de recherche INRIA Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex (France)

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399