An Analysis of Some Models Used in Image Segmentation

Robin Morris, Xavier Descombes, Josiane Zerubia

INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

An analysis of some models used in image segmentation

Robin Morris, Xavier Descombes and Josiane Zerubia

N° 3016

October 1996

THÈME 3

ISSN 0249-6399

INRIA
SOPHIA ANTIPOLIS

An analysis of some models used in image segmentation

Robin Morris, Xavier Descombes and Josiane Zerubia

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

Rapport de recherche n ° 3016 — October 1996 — 30 pages

Abstract: This report describes an investigation into the characteristics of a number of image models and algorithms used for the segmentation of SPOT images. The initial goals were the problems of phase transition in the model, and parameter estimation. The experimental tools used encompass a great deal of modern Markov chain Monte Carlo (MCMC) methodology, from the Gibbs sampler/Metropolis-Hastings algorithm, to the Swendsen-Wang algorithm and Monte Carlo Maximum Likelihood. One result of this work has been the demonstration of the importance of in-depth studies of the image models being considered – some common models are shown to be inadequate for the purposes to which they are commonly put. Outlines of future areas of research aimed at overcoming some of the problems identified are given.

Key-words: Image segmentation, Markov Random Fields, Hierarchical model, Parameter Estimation, Markov chain Monte Carlo, Swendsen-Wang algorithm

(Résumé : tsvp)

* R.D. Morris was supported by a grant from the Commission of the European Communities under the HCM program.
** email: *name*@sophia.inria.fr

Unité de recherche INRIA Sophia-Antipolis
2004 route des Lucioles, BP 93, 06902 SOPHIA-ANTIPOLIS Cedex (France)
Téléphone : (33) 93 65 77 77 – Télécopie : (33) 93 65 77 65

Analyse de quelques modèles utilisés en segmentation

Résumé : Ce rapport présente un travail mené en vue de l'étude des caractéristiques d'un certain nombre de modèles et d'algorithmes utilisés pour la segmentation d'images SPOT. Initialement cette étude concernait les problèmes de transition de phases et d'estimation de paramètres. Les méthodes employées incluent un grand nombre de méthodes modernes fondées sur un principe de type "MCMC": l'échantillonneur de Gibbs et l'algorithme de Metropolis-Hasting, l'algorithme de Swendsen Wang et l'estimation au sens du maximum de vraisemblance par une méthode de Monte Carlo. Un des résultats de ce travail a été la mise en évidence de l'importance d'une étude approfondie des modèles considérés. Certains modèles utilisés s'avèrent inadaptés au type de problèmes qu'ils sont censés résoudre. Des ébauches d'axes de recherche futurs visant à surmonter certains de ces problèmes sont exposées.

Mots-clé : segmentation d'image, champs de Markov, modèle hiérarchique, estimation de paramètres, MCMC, algorithme de Swendsen-Wang.

Contents

1 Introduction

This report concerns methods for the segmentation of satellite imagery, the goal being to assign to each pixel of the image a label, indicating to which class the pixel belongs [3, 14]. This is of interest in, for example, land use management, where people are interested in the spatial distribution of different crop types.

We begin in section 2 by describing the monoscale and hierarchical models which have been used for segmentation purposes [5, 18, 12, 13]. The first of these is based on the use of the Ising/Potts model as a regulariser; the second constructs a mutiresolution pyramid with cliques within and between the resolution levels. The within-level clique potentials are derived from the original monoresolution model.

Both of these models have two types of parameters associated with them – the *class parameters* and the *model parameters* or hyperparameters. The class parameters describe what we expect to observe at a pixel, given its class membership. The model parameters are intended to adjust the behaviour of the model such that realisations from the model are representative of the segmented images we desire to achieve. The two types of parameter have clearly distinct roles and are usually treated separately [2]. In this report we focus mainly on the segmentation models themselves, and are thus concerned primarily with the model parameters.

The remainder of the report is an investigation into the characteristics of the Potts and hierarchical models as applied to image segmentation.

The main tools used are Markov chain Monte Carlo (MCMC) methods [17, 21, 23]. The models are probabilistic and analytically intractable. MCMC methods are based on the simulation of these distributions by a Markov chain which realises samples from the distribution. From these samples we may draw inferences about functions with respect to the distribution.

For the Potts model it is known, and for the hierarchical model we postulate, (see section 6), that the model undergoes a phase transition [6]. The presence of a phase transition has a number of effects – the characteristics of realizations of the distribution change suddenly as the model parameters vary. It also causes the effect of 'critical slowing down' in many MCMC simulation algorithms. This causes the realizations of even long runs of single site updating algorithms to not be truly representative of the distribution of interest. The Swendsen-Wang algorithm [22, 10] was designed to overcome these problems and provides an ingenious algorithm which converges rapidly even near the phase transition. (It is also used as part of recent algorithms for *exact* simulation [19]) The Swendsen-Wang algorithm is detailed in section 3 and used extensively in the experimental work in the remainder of this report.

The difficulty in estimating the model parameters is due to the analytic intractability of the normalising constant (or partition function) associated with the models' probability density function. An early work-around was to use the pseudolikelihood [1].

In section 4 an approach to parameter estimation is detailed. In the case of fully observed (ie noise free) data, it is shown that the parameters can be estimated by matching the mean value of the natural statistics observed in a number of samples with the values of the natural

statistics of the observed data. This allows limited parameter estimation to be performed using look up tables. It also results in a powerful mechanism for model criticism – when model parameters can truly be estimated we can begin to decide whether realisations of the model are actually representative of the segmentations we are trying to achieve.

The approach to parameter estimation in section 4 does still require a great deal of computation, and is inflexible in that the simulations must be performed for each image size and number of classes. In section 5 we detail the method of MCMC Maximum Likelihood [7]. This allows parameter estimation to be performed more economically, by re-using the samples from one set of parameter values when calculating the partition function (and its derivatives) at other values.

In section 6 we touch on the issue of phase transitions. It is known that in physical systems a phase transition is indicated by a discontinuity in the 'specific heat' [8]. Considering the models as imaginary physical systems, the specific heat may be estimated from the samples. We present results showing the probable presence of a phase transition in the hierarchical model.

Finally, we present some results of performing segmentations with the models discussed. We demonstrate that the MAP estimates using the parameters estimated from the known segmentations do not correspond to the desired results. This indicates that the models are not capturing the desired characteristics of the ideal segmentations.

Section 8 concludes the report.

2 The monoscale and hierarchical models for segmentation

Image segmentation is an ill-posed problem, that is, the data in themselves are insufficient to unambiguously define the segmentation [20]. In this case it is necessary to introduce some additional information. In its most general form, this is often that '*neighbouring pixels tend to be of the same class*', *i.e.* , the segmentation should be *smooth* in some sense. This leads naturally to the use of MRFs as the priors on the desired segmentation. These models are attractive from a computational point of view, as the global properties derive solely from local interactions.

2.1 The monoscale model

This is the simplest form of MRF used in image segmentation. The model for the segmented image is the Ising (in the case of binary segmentations) or Potts (for the general N-colour) model of statistical physics. The probability of the segmentation taking a configuration $\mathbf{x} \in \Omega$ is given by

$$p(\mathbf{x}|\beta) = \frac{1}{Z(\beta)} \exp\left(\beta \sum_{i \sim j} \delta(x_i - x_j)\right) \tag{1}$$

where $i \sim j$ indicates that the summation is over neighbouring pairs. Traditionally the 4 nearest-neighbours or 8 nearest-neighbours are used. This distribution may be written more instructively as

$$p(\mathbf{x}|\beta) = \frac{1}{Z(\beta)} \exp\left(\beta \mathrm{NHC}(\mathbf{x})\right) \tag{2}$$

where $\mathrm{NHC}(\mathbf{x})$ is the number of homogeneous cliques in the realisation $\mathbf{x}$. This shows clearly that the probability of a realisation is a global function of the realisation.

The second component of a segmentation model is to define the class conditional probabilities, *i.e.* given that a pixel is of a particular class, what is the probability of it taking each of the observable values? Conditional on the segmentation, this distribution at each pixel is usually taken to be independent of its neighbours. A commonly used model is that the observed value, f_i, is drawn from a gaussian distribution, $\mathcal{N}(\mu_{x_i}, \sigma_{x_i})$ where μ_{x_i} and σ_{x_i} are the mean and variance associated with the class x at pixel i. Thus, the posterior distribution for the segmentation is

$$\begin{aligned} p(\mathbf{x}|\mathbf{f}, \beta, \{\sigma, \mu\}) \propto & \frac{1}{Z(\beta)} \exp(\beta \mathrm{NHC}(\mathbf{x})) p(\beta) p(\{\mu, \sigma\}) \\ & \times \prod_i \frac{1}{\sqrt{2\pi}\sigma_{x_i}} \exp\left(\frac{(f_i - \mu_{x_i})^2}{2\sigma_{x_i}}\right) \end{aligned} \tag{3}$$

where $\mathbf{f}$ is the observed data. Commonly the priors $p(\beta)$ and $p(\{\mu, \sigma\})$ are taken to be uniform and hence ignored.

Once the posterior has been defined, the segmentation is performed by minimising a cost function. Two common cost functions result in the maximum a-posteriori (MAP) solution, found using simulated annealing (SA), or, more naturally in the case of image segmentation, the maximum of the posterior marginals solution (MPM) [15], which minimises the expected number of misclassified pixels, found by using a Markov chain Monte Carlo (MCMC) algorithm.

2.2 The hierarchical model

In an attempt to introduce longer-range dependencies, to improve the resulting segmentation when used with sub-optimal deterministic algorithms, and to allow more efficient implementation on parallel architectures, the hierarchical model discussed in this section was introduced in [12]. This is a development of the *multiscale* algorithm in [18].

The aim is to define a multiscale pyramid, where at subsequent levels the size is reduced, usually by a factor of two. At each level, the aim is to have the same energy function as in the original monoscale formulation. Links between the levels are then introduced to try to induce longer range dependencies. Figure 1 illustrates the pyramid and the within and between level neighbourhood structure.

Consider the n^{th} $(n > 0)$ level of the pyramid, where level 0 is the original level. This consists of *blocks* of size $l^n \times l^n$, where l is the subsampling factor. Typically $l = 2$. To define

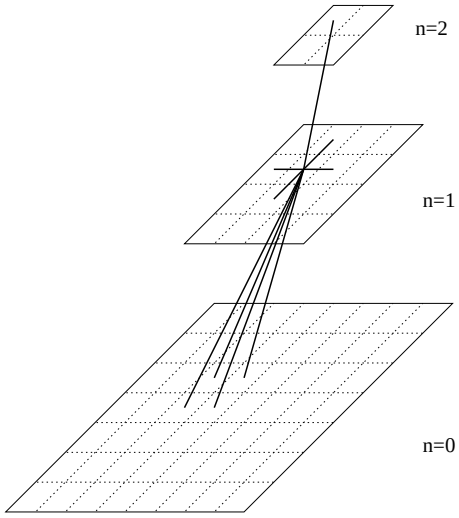

Figure 1: Neighbourhood structure on the pyramid

the energy function at level n in terms of the level n blocks, it is necessary to consider how the cliques at level 0 are mapped into the structure at level n. In this study we consider only pairwise interactions at level zero[1] An adjacent pair of pixels at level 0 may fall into one of two categories at level n:

1. both pixels may fall *within* a single block
2. the pixels may fall on into each of two neighbouring blocks

At level n the blocks x_i^n are allocated a single class. Thus, all cliques falling into category 1 contribute $+\beta$ to the energy function at level n. There are

$$p_n = 2l^n(l^n - 1)$$

such cliques in each block. Along the edge of each block there are

$$q_n = l^n$$

cliques at the original level. Thus the energy function at level n is

$$U_n(\mathbf{x}^n) = \beta(q_n \mathrm{NHC}(\mathbf{x}^n)) + \beta p_n \ \mathrm{NB}(n) \tag{4}$$

where $\mathrm{NB}(n)$ is the number of blocks at level n (for $n > 0$).

[1] The general case is described in [18].

At each level the observed image data is also included in a similar manner to the mono-scale model. At level n the blocks are expanded back to the original scale, and then same independent gaussian model for the class conditional probabilities are used.

The difference between the hierarchical model and the multiscale model is the following: in the hierarchical model we now introduce cliques between the levels. Each block at level n has one parent and l^2 children. Each such clique contributes γ to the global energy of the entire pyramid if it is homogeneous (zero otherwise) *irrespective* of which level it is situated on. Thus the posterior distribution for the segmentation pyramid, assuming uniform priors on the parameters, is

$$p(\mathbf{x}|\mathbf{f},\beta,\gamma,\{\mu,\sigma\}) = \frac{1}{Z(\beta,\gamma)} \exp\left(\sum_{n=0}^{N} [\beta q_n \mathrm{NHC}(\mathbf{x}^n) + \beta p_n \mathrm{NB}(n)] + \gamma \mathrm{NHCBL}(\mathbf{x})\right) \times \prod_{n=0}^{N} \prod_{i} \frac{1}{\sqrt{2\pi}\sigma_{x_i^n}} \exp\left(\frac{(f_s - \mu_{x_i^n})^2}{2\sigma_{x_i^n}}\right)$$

Where $\mathrm{NHCBL}(\mathbf{x})$ is the number of homogeneous cliques between the levels of the entire pyramid. The first term is thus the prior on the pyramid, and the second term indicates that at each level, the pyramid is expanded out to the original resolution, and the correspondence with the data is made at that resolution.

3 The Swendsen-Wang algorithm

Due to the structure of the models discussed above analytic computations are impossible, and they are instead simulated using MCMC algorithms. The most well known of these are the Metropolis-Hastings (M-H) algorithm [16, 9] and a special case, the Gibbs sampler [5]. In their usual implementations these algorithms update a single pixel (or element of the pyramid) at a time. It has become more widely appreciated recently that single site updating algorithms can take a very long time to converge, such that the realisation is truly representative of the distribution being studied [11].

However, for simulating the Potts model the Swendsen-Wang (SW) algorithm was developed, which converges rapidly to the equilibrium distribution and then moves relatively freely within this distribution. It is an example of an auxiliary variable algorithm. In this section we will describe the SW algorithm for use when simulating the Potts model without data (the extension for simulating the hierarchical model is straightforward). the SW algorithm in its basic form is less useful when studying models with data; in [10] modifications have been introduced to help ameliorate this.

The idea behind auxiliary variable methods is the following:

It is desired to simulate a distribution $\pi(\mathbf{x})$. Auxiliary variables $\mathbf{u}$ are introduced, with conditional distribution $\pi(\mathbf{u}|\mathbf{x})$. This gives a joint distribution $\pi(\mathbf{x},\mathbf{u}) = \pi(\mathbf{u}|\mathbf{x})\pi(\mathbf{x})$, with the desired marginal distribution for $\mathbf{x}$ of $\pi(\mathbf{x})$. Simulation of this distribution is generally

performed by alternately updating $\mathbf{u}$ and $\mathbf{x}$ – the idea being to define $\pi(\mathbf{u}|\mathbf{x})$ such that the updates cause rapid mixing. The realisations of $\mathbf{x}$ are those desired.

For the SW algorithm the distribution $\pi(\mathbf{u}|\mathbf{x})$ is defined such that the u_{ij} are independent, and is

$$p(u_{ij}|\mathbf{x}) \propto \exp\left(-\beta \mathrm{I}[x_i = x_j]\right) \mathrm{I}[0 \leq u_{ij} \leq \exp\left(\beta \mathrm{I}[x_i = x_j]\right)] \tag{5}$$

where u_{ij} can be considered as a continuous 'bond' variable between the pixels x_i and x_j and $\mathrm{I}[\cdot]$ is the indicator function. This results in

$$\pi(\mathbf{x}|\mathbf{u}) \propto \prod_{i \sim j} \mathrm{I}[0 \leq u_{ij} \leq \exp\left(\beta \mathrm{I}[x_i = x_j]\right)] \tag{6}$$

What does this choice of distribution give us? Considering first $p(u_{ij}|\mathbf{x})$, for $u_{ij} > 1$ we must have $\exp\left(\beta \mathrm{I}[x_i = x_j]\right) > 1$, or equivalently, $x_i = x_j$. Thus $u_{ij} > 1$ constrains x_i and x_j to be in the same state. Conversely, if x_i and x_j are in the same state, what is the probability of $u_{ij} > 1$? From the conditional distribution in equation 5 we have that

$$p(u_{ij} > 1 | x_i = x_j) = 1 - \exp(-\beta) \tag{7}$$

Since it is only important whether u_{ij} is greater or less than one we may think of the u_{ij} as binary bond variables. From equation 7 the bond variable is present between two pixels in the same state with probability $1 - \exp(-\beta)$. To sample $\mathbf{u}$ thus involves placing bonds between neighbouring pixels of the same state with probability $1 - \exp(-\beta)$ and omitting bonds between neighbouring pixels of differing states.

Once the bonds are in place, the conditional distribution $\pi(\mathbf{x}|\mathbf{u})$ says that all configurations where bonded pixels are of the same state are equally probable. Thus to update $\mathbf{x}$ we form clusters of connected pixels and assign to all pixels of the cluster the same state, chosen uniformly from the allowed states. This scheme allows potentially large clusters of pixels to change state at each iteration, allowing the Markov chain to explore the distribution freely.

For small β the links will only rarely be placed, resulting in a large number of small clusters. In this case standard single-site updating algorithms are more efficient. For large β (and especially β close to the critical value (see section 6)) the improved mobility of the SW algorithm more than compensates for the extra computation involved.

3.1 The Swendsen-Wang algorithm for the hierarchical model

To implement the SW algorithm for the hierarchical model is straightforward. Within each level the bonds are placed with probability $1 - \exp(-\beta_n)$, where $\beta_n = \beta q_n$, and between levels the bonds are placed with probability $1 - \exp(-\gamma)$. This forms 3 dimensional clusters on the pyramid, which are then coloured randomly.

3.2 Realisations of the Potts and Hierarchical models

Here we present some simulations of the Potts model for various values of β (figure 3) and the hierarchical model (figure 4) for various values of β and γ.

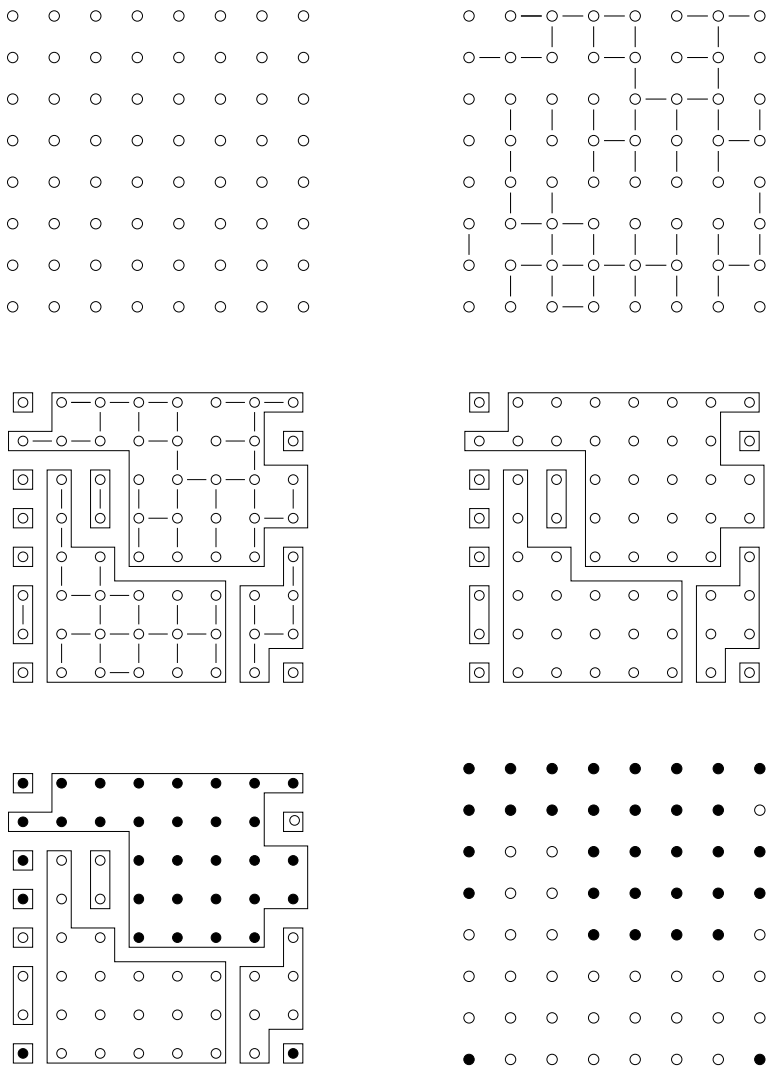

Figure 2: Starting from a uniform image (top left), links are placed independently between the pixels with probability $1-\exp(-\beta)$ (here $\beta = 0.7$) (top right). Clusters are formed using the links (middle left). Once the clusters are in place, the links can be forgotten (middle right). All the pixels in a cluster are coloured with the same randomly chosen colour (bottom left). The final image shows that in a single update, very many pixels can change state

Note that for the Potts model the realisations do not show the uniform patches that this model is often claimed to produce – these are artefacts of not running a single site updating algorithm long enough. For the hierarchical model, note that the higher levels of the pyramid rapidly converge to one colour as β is raised. This is due to the scaling of β to produce β_n – even for small values of β, β_n rapidly takes on large values as n increases. The parameter γ controls the influence successive levels have on each other. As γ is increased levels rapidly begin to closely resemble each other. This tight linking between different elements of the pyramid causes single site updating algorithms to become stuck very easily. The final image for the hierarchical model shows that it is possible to choose parameters which give some order at the highest level, but this is at the expense of having almost totally random behaviour at lower levels, which will not provide enough smoothing for good segmentations.

In the remainder of this report, we use exclusively realisations generated by the SW algorithm, because of its superior convergence and mobility for the models we are considering. We can thus be more certain that the realisations we use are truly representative of the distributions of interest.

We now begin to address the problem of parameter estimation for the two models being considered. Recall that we are interested here in the model parameters rather than the class conditional parameters, that is the parameter β for the Potts model and the parameters β and γ for the hierarchical model.

4 Parameter estimation for fully observed data

In this section we demonstrate that for fully observed data (*i.e.* in the current context, realisations of the label field) MCMC methods allow the model parameters to be estimated accurately.

4.1 The Potts model

Recall that the Potts model is given by equation 1, repeated here

$$p(\mathbf{x}|\beta) = \frac{1}{Z(\beta)} \exp\left(\beta \sum_{i \sim j} \delta(x_i - x_j)\right) \tag{8}$$

The difficulty in estimating the parameter β given $\mathbf{x}$ (either using Maximum Likelihood (ML) or a Bayesian estimator) is due to the normaliser of the distribution. This is given by

$$Z(\beta) = \sum_{\{x\}} \exp(\beta \times \mathrm{NHC}(\mathbf{x})) \tag{9}$$

where the summation is over all possible segmentations – an intractably large number.

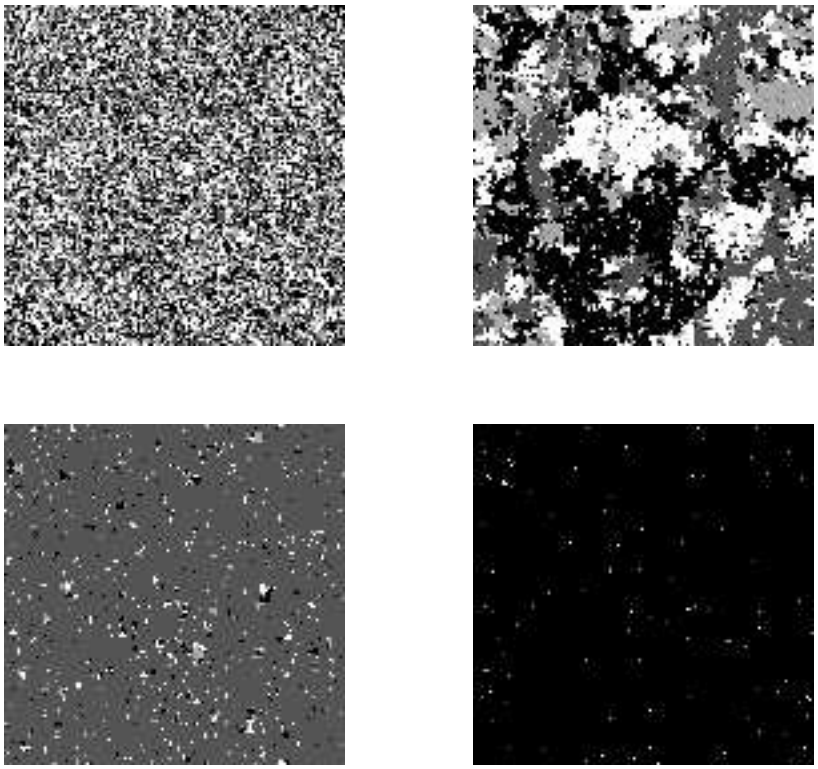

Figure 3: Realisatons of the 4-state Potts model for $\beta = 0.2, 0.5493, 0.6, 0.8$ (top left, top right, bottom left, bottom right)

Given an image $\hat{\mathbf{x}}$, the likelihood for β is

$$p(\beta|\hat{\mathbf{x}}) = \frac{1}{Z(\beta)} \exp(\beta \times \mathrm{NHC}(\hat{\mathbf{x}}))$$

where

$$Z(\beta) = \sum_{\{x\}} \exp(\beta \times \mathrm{NHC}(\mathbf{x}))$$

The maximum likelihood estimate is given by

$$\frac{\partial \log p(\beta|\hat{\mathbf{x}})}{\partial \beta} = \mathrm{NHC}(\hat{\mathbf{x}}) - \frac{Z'(\beta)}{Z(\beta)} = 0$$

Now

$$\frac{Z'(\beta)}{Z(\beta)} = \frac{\sum_{\{x\}} \mathrm{NHC}(\mathbf{x}) \exp(\beta \times \mathrm{NHC}(\mathbf{x}))}{Z(\beta)}$$

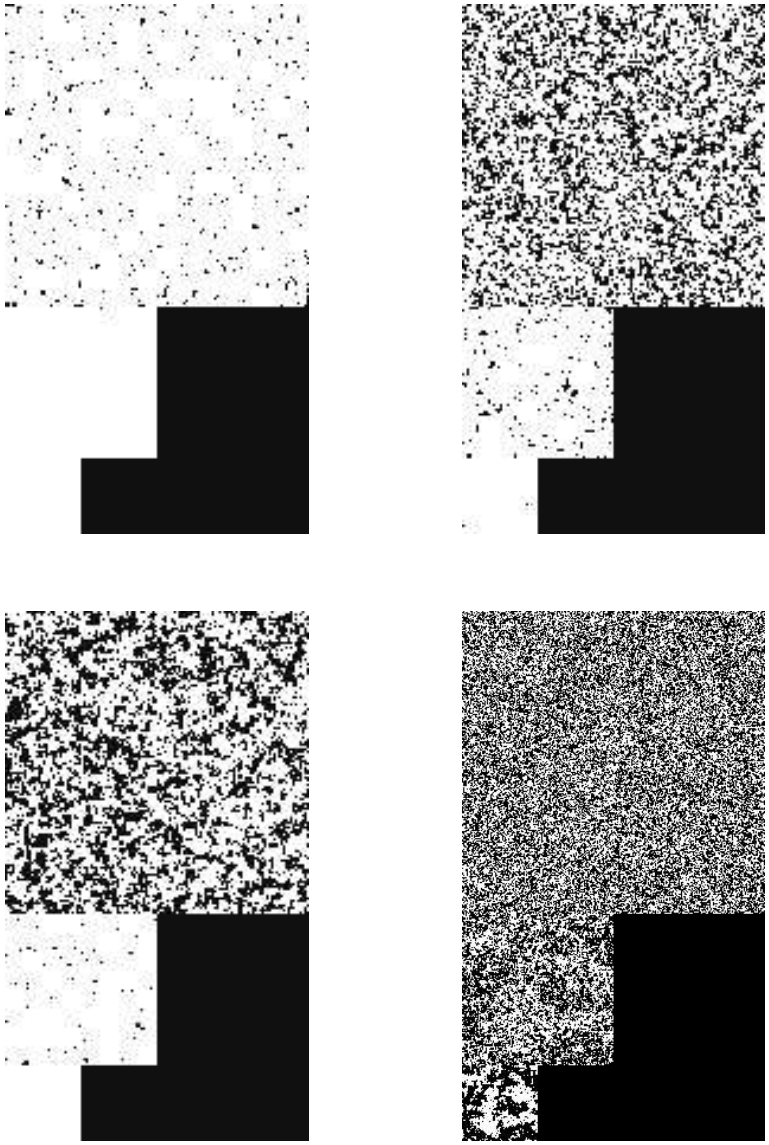

Figure 4: Realisations of the hierarchical model for (top row) $\beta = 0.5$, $\gamma = 0.1$ and $\beta = 0.2$, $\gamma = 0.15$ (bottom row) $\beta = 0.3$, $\gamma = 0.03$ and $\beta = 0.1$, $\gamma = 0.11$

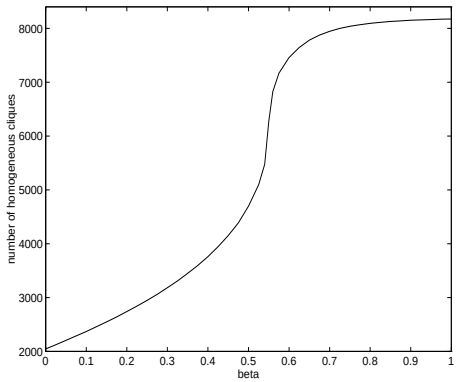

Figure 5: $< N(\mathbf{x}) >$ vs β for $64 \times 64 \times 4$-state Potts model

$$\begin{aligned} &= \sum_{\{x\}} \mathrm{NHC}(\mathbf{x}) p(\mathbf{x}|\beta) \\ &= < \mathrm{NHC}(\mathbf{x}) >_\beta \end{aligned}$$

where $< \mathrm{NHC}(\mathbf{x}) >_\beta$ is the mean number of homogeneous cliques with respect to the distribution $p(\mathbf{x}|\beta)$ and hence the maximum likelihood estimate of β is the value of β for which

$$\mathrm{NHC}(\hat{\mathbf{x}}) = < \mathrm{NHC}(\mathbf{x}) >_\beta$$

Note that $< \mathrm{NHC}(\mathbf{x}) >_\beta$ is a function of β and that analytically it is intractable. However an estimate of the number of homogeneous cliques under the distribution defined by a particular value of β can be easily performed by generating samples from $p(\mathbf{x}|\beta)$ and forming an empirical average.

Figure 5 shows $< N(\mathbf{x}) >_\beta$ as a function of β for a 4-state Potts model on a 64×64 lattice. The graph is the empirical mean number of homogeneous cliques for 1400 samples generated by the Swendsen-Wang algorithm [22], for 43 values of β between 0 and 1. The graph traces from $\beta = 0$, corresponding to a completely random image (no correlation between neighbouring pixels) to $\beta = 1$, where the image is almost entirely one colour (indicated by the image having a number of homogeneous cliques very close to the total number of cliques in the image). Estimates for β from fully observed data may be made simply by calculating $N(\hat{\mathbf{x}})$ and reading the corresponding β from the graph.

Figure 6 shows two image used to test segmentation algorithms. The first has 7619 homogeneous cliques, corresponding to $\beta \simeq 0.625$. The second has 5769 homogeneous cliques, corresponding to $\beta \simeq 0.544$. The first of these is a value of β which gives a very uniform image, rather than one comprised of zones, and the second gives a fairly random image, with zones of many sizes, providing insufficient smoothing for a successful segmentation (see figure 3).

Figure 6: Test Images

4.2 The hierarchical model

The parameters β and γ of the hierarchical model may be estimated from realisations of the segmentation pyramid in a similar manner. Recall that the energy function for the hierarchical model is

$$\begin{aligned} U(\mathbf{x}) &= \sum_{n=0\ldots N} \beta q_n \mathrm{NHC}(\mathbf{x}^n) + \beta p_n \mathrm{NB}(n) \\ &\quad + \gamma \mathrm{NHCBL}(\mathbf{x}) \end{aligned} \tag{10}$$

and is governed by two statistics. The first we term the *effective number* of homogeneous within level cliques, and is

$$N_1(\mathbf{x}) = \sum_{n=0\ldots N} q_n \mathrm{NHC}(\mathbf{x}^n) + p_n \mathrm{NB}(n) \tag{11}$$

and the second is the number of homogeneous between-level cliques,

$$N_2(\mathbf{x}) = \mathrm{NHCBL}(\mathbf{x}) \tag{12}$$

In a similar manner to the simulations used to generate figure 5, realisations of the hierarchical model were generated for a range of values of β and γ, for a 2-state model with 3 levels, a resolution of 64×64 pixels at level 0, and $l = 2$. Figure 7 shows the contours of $N_1(\mathbf{x})$ and $N_2(\mathbf{x})$ as functions of β and γ.

Figure 8 shows two test images, and the corresponding sections of figure 7 which enable the parameters to be estimated. The first figure has $N_1(\mathbf{x}) = 22802$, $N_2(\mathbf{x}) = 5062$. The corresponding parameters are found where the central dotted line crosses the central solid line. This is somewhere off the top left of the figure, indicating a very low value of β and a very high value of γ. This is consistent with the images in figure 8 – because the squares correspond with the subsampling to form the pyramid, we require almost all the cliques

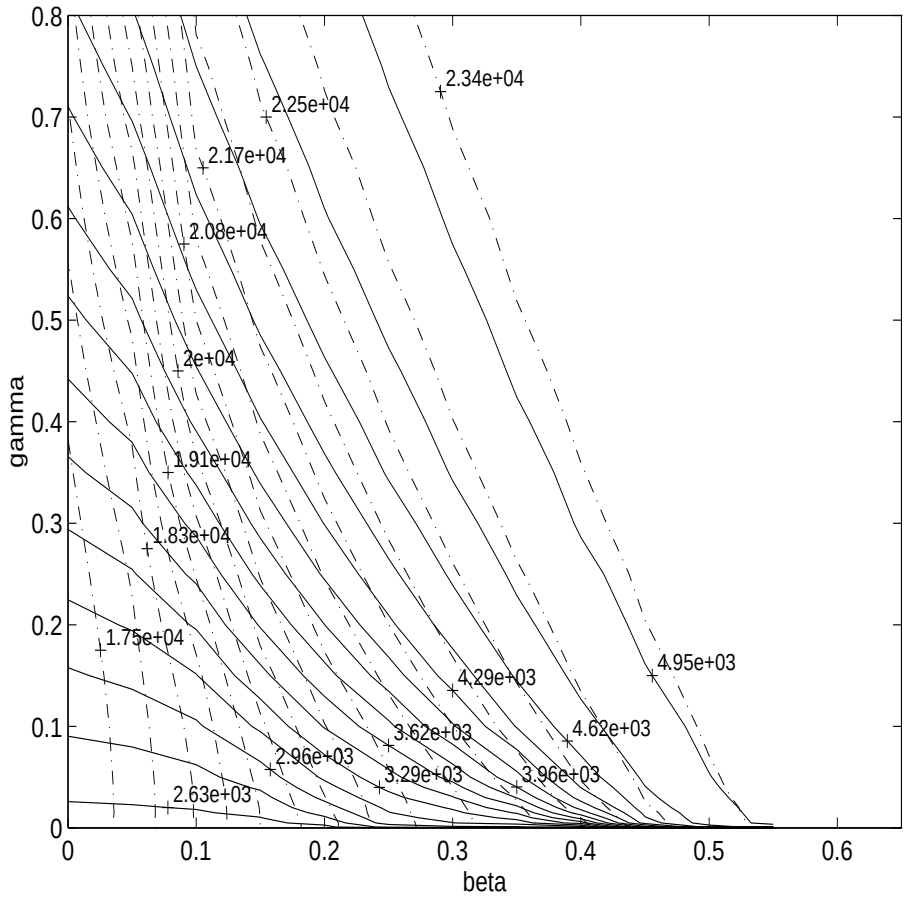

Figure 7: $N_1(\mathbf{x})$ (dash-dot) and $N_2(\mathbf{x})$ (solid) as functions of β and γ (see text)

between the levels to be homogeneous, resulting in a high value for γ. With a high value of γ, at the lower levels we will have very many homogeneous cliques, due to the spatial correlation induced by the pyramid structure, even with a very small value of β. The second image has $N_1(\mathbf{x}) = 22484$, $N_2(\mathbf{x}) = 4721$, corresponding to $\beta = 0.15, \gamma = 0.6$ – because the squares no longer correspond with the subsampling, the parameters take on less extreme values.

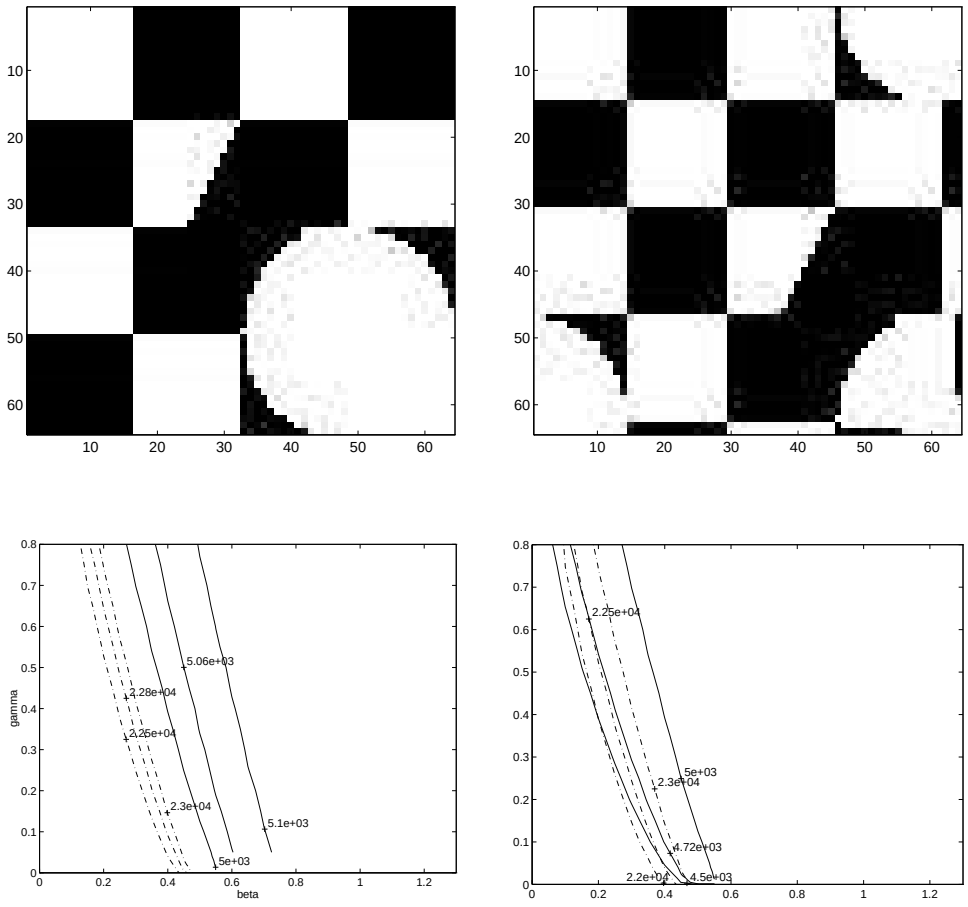

Figure 8: Test images and graphs for parameter estimation for the hierarchical model

5 MCMCML

In the previous section we showed that it is possible to do approximate maximum likelihood estimation for the parameters of the models that we are interested in. However the method discussed in section 4 whilst being of interest for obtaining a global perspective of the parameter estimation problem, is not really practical for rapid on-line estimation.

The method of Markov chain Monte Carlo Maximum Likelihood (MCMCML, or sometimes just Monte Carlo Maximum Likelihood, MCML) addresses this problem, and allows practical algorithms for estimation to be constructed. It is based on the principle of importance sampling – the idea that we can use samples drawn from one distribution to estimate expectations with respect to another distribution, by suitable reweighting. What this allows us to do here is that we can use samples drawn from $\{\beta_0, \gamma_0\}$ to calculate the likelihood and its gradients for parameters $\{\beta_1, \gamma_1\}$. Thus we can *re-use* the samples we have generated, leading to more efficient algorithms. In theory, we can use the samples from $\{\beta_0, \gamma_0\}$ to calculate expectations for *any* $\{\beta_1, \gamma_1\}$; in practice, if the distributions do not overlap significantly the importance sampling weights will be effectively zero apart from the sample which has the highest probability under $\{\beta_1, \gamma_1\}$, which will have a weight of one. This will lead to extremely inaccurate estimates, and in practice we must re-sample when $\{\beta, \gamma\}$ have changed significantly.

In general, the models we are considering (together with many others) can be written in the form

$$p(\mathbf{x}|\theta) = \frac{1}{Z(\theta)} \exp\left(\sum_{i=1\ldots K} \theta_i N_i(\mathbf{x}) \right)$$

where there are K statistics N_i of interest. The negative log likelihood, which we wish to minimise, is

$$-l(\theta; \mathbf{x}) = \log(Z(\theta)) - \theta_1 N_1(\mathbf{x}) - \theta_2 N_2(\mathbf{x}) + \ldots$$

and the gradients are

$$\frac{\partial(-l(\theta; \mathbf{x}))}{\partial \theta_i} = < N_i >_\theta - N_i(\mathbf{x})$$

Also, but less usefully as the estimation accuracy is often poor, the elements of the Hessian are

$$\frac{\partial}{\partial \theta_i \partial \theta_j} - l(\theta; x) = < N_i >_\theta < N_j >_\theta - < N_i N_j >_\theta$$

So, to perform efficient Maximum Likelihood estimation for this class of models we require to be able to estimate $Z(\theta)$ and $< N_i >_\theta$ for a wide range of values of θ, using samples generated with the minimum possible number of values of θ.

Given samples $\mathbf{x}^j$ generated from $p(\mathbf{x}|\phi)$, we may calculate the required quantities in the following manner.

$$Z(\theta) \quad = \quad \sum_{\{x\}} \exp\left(\sum_i \theta_i N_i(\mathbf{x}) \right)$$

$$
\begin{aligned}
&= \sum_{\{x\}} \exp\left(\sum_i (\theta_i - \phi_i) N_i(\mathbf{x})\right) Z(\phi) p(\phi) \\
&= Z(\phi) < \exp\left(\sum_i (\theta_i - \phi_i) N_i(\mathbf{x})\right) >_\phi \\
\frac{Z(\theta)}{Z(\phi)} &\simeq \frac{1}{N} \sum_{j=1\ldots N} \exp\left(\sum_i (\theta_i - \phi_i) N_i(\mathbf{x}^j)\right) \qquad (13)
\end{aligned}
$$

And

$$
\begin{aligned}
< N_i >_\theta &= \sum_{\{x\}} N_i(\mathbf{x}) p(\mathbf{x}|\theta) \\
&= \sum_{\{x\}} \frac{N_i(\mathbf{x})}{Z(\theta)} \exp\left(\sum_i \theta_i N_i(\mathbf{x})\right) \\
&= \frac{Z(\phi)}{Z(\theta)} \sum_{\{x\}} N_x(\mathbf{x}) \exp\left(\sum_i (\theta_i - \phi_i) N_i(\mathbf{x})\right) p(\mathbf{x}|\phi) \\
&\simeq \frac{Z(\phi)}{Z(\theta)} \frac{1}{N} \sum_j N_i(\mathbf{x}^j) \exp\left(\sum_i (\theta_i - \phi_i) N_i(\mathbf{x}^j)\right)
\end{aligned}
$$

And any minimisation procedure (gradient descent, golden section search, conjugate gradient, etc) can be used with these values to effect the minimisation. As mentioned above, when θ is far from ϕ (in some appropriate sense), the estimation will be inaccurate. So practically, we must define a criteria for deciding when it is necessary to resample with a new value of ϕ. This can be done by monitoring the *importance sampling weights*, which in this case are $\exp(\sum_i (\theta_i - \phi_i) N_i(\mathbf{x}^j))$. When the range of these weights becomes large, it is necessary to resample.

Figure 9 shows a binary test image. As discussed earlier, this type of image causes the parameters of the hierarchical model to take on extreme values. Also shown is a sample from the hierarchical model with the estimated parameters. This is almost totally one colour, with small points of the other colour. Clearly this will not provide a suitable degree of smoothing for a segmentation (see section 7).

Figure 10 shows a section from a SPOT image segmented into five classes. Also shown are realisations of the Potts and hierarchical model with parameters estimated from this image, and a realisation of the *chien model* [4] with parameters estimated from the segmented SPOT image. The hierarchical and Potts models clearly do not provide suitable smoothing. The chien model does seem to have captured better the structure in the original image.

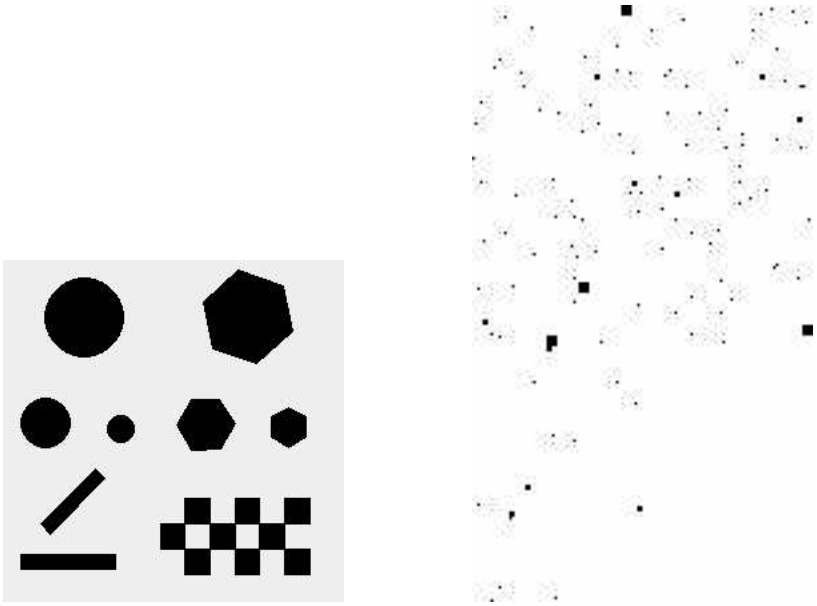

Figure 9: Test image and resynthesised hierarchical model – $\beta = 0.0684, \gamma = 2.36$

6 Phase Transitions

The presence of a phase transition in the behaviour of the model can cause difficulties when generating the samples used in the parameter estimation schemes described above. It is well known that the Potts model undergoes a phase transition at $\beta_c = 0.5\log(1+\sqrt{n})$ (n-colours) [6] – that below this value the realisations are fairly random, close to this value order rapidly emerges and above it the realisations are rapidly dominated by one colour. It is also known that close to β_c single site updating algorithms such as the usual forms of the Metropolis-Hastings algorithm or the Gibbs sampler suffer from what is termed *critical slowing down* – that they explore the state space extremely slowly. This slow exploration of the state space is a practical problem – if the samples we use to estimate the parameters are not truly representative of the distribution then the estimates will be inaccurate.

The Swendsen-Wang algorithm goes some way to alleviate the situation, but it is still necessary to take great care to ensure that the simulations have indeed converged, and have truly explored the distribution.

Physically, the presence of a phase transition is indicated by a discontinuity in the *specific heat* [8]. The specific heat can be estimated from realisations of the model, by considering the variance of the energy

$$\hat{c}_T \propto \frac{1}{K}\sum_{i=1}^{K}\left[U(\mathbf{x}^i) - \bar{U}(\mathbf{x})\right]^2$$

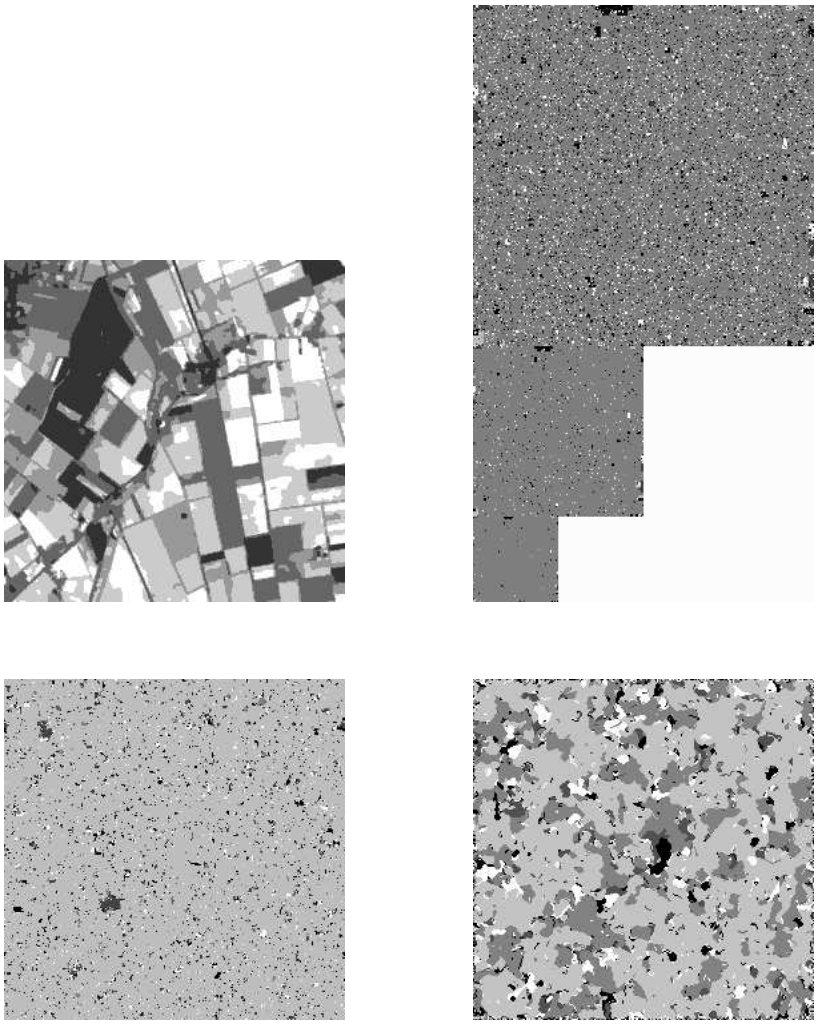

Figure 10: Segmented SPOT image. Resynthesised hierarchical model with estimated parameters ($\beta = 0.08, \gamma = 1.32$), Resynthesised Potts model ($\beta = 0.6043$) and the corresponding *chien model*

Figure 11 shows the variance of the energy of a 128×128 Ising model. The phase transition is clear. Figure 12 shows the variance of the energy of the hierarchical model as a function of β and γ for a pyramid with three levels, size 64×64 at the bottom level and two colours. Because of the manner in which β changes at different levels in the pyramid the phase

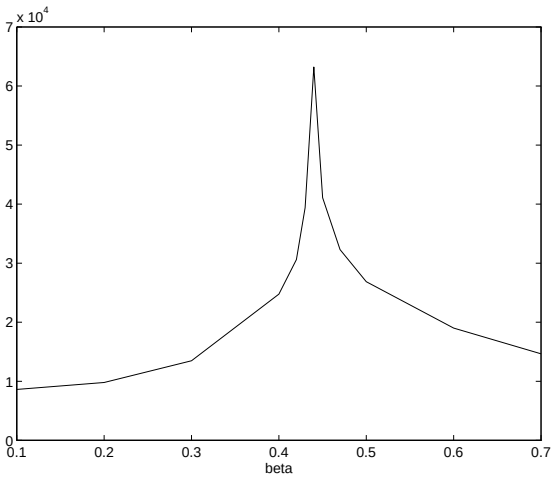

Figure 11: Variance of Energy, Ising model

transition is not as strong as for the Ising model (also the lattice considered is smaller), but is clearly present, tracing from near β_c for the Ising model at $\gamma = 0$.

7 Segmentation Results

In this section we present results of segmentations of images degraded by both channel noise (binary images) and white noise (grey scale images). We compare the results obtained using the parameters estimated from the noise-free images with those obtained by using parameters chosen 'by an expert', which in essence means that numerous different parameter values were chosen by intuition, and the 'best' resulting segmentation was retained.

The results for the binary images corrupted by channel noise are given in figures 13 and 14. The results with the estimated parameters show that the characteristics of the image captured by the model are not those desired – while the results might be optimal in terms of the MAP criterion, they are deficient in terms of the 'users' criterion. The model does not correspond to the prior information the user would wish to impose on the solution. Figure 13 shows that the model can cope when the degradation is not too strong, but figure 14 shows that for increased noise, when the regularising effect of the model is more important, that the model does not provide the desired smoothing becomes clear.

Figures 15 and 16 have the same interpretation. When the noise level is increased the effect of the prior becomes more important, and it is then that we see that the model does not bring the desired characteristics to the results. By choosing parameters by hand the segmentation can be made more similar to what the user would desire, but with these parameter values the model does not really correspond to the original image. The need

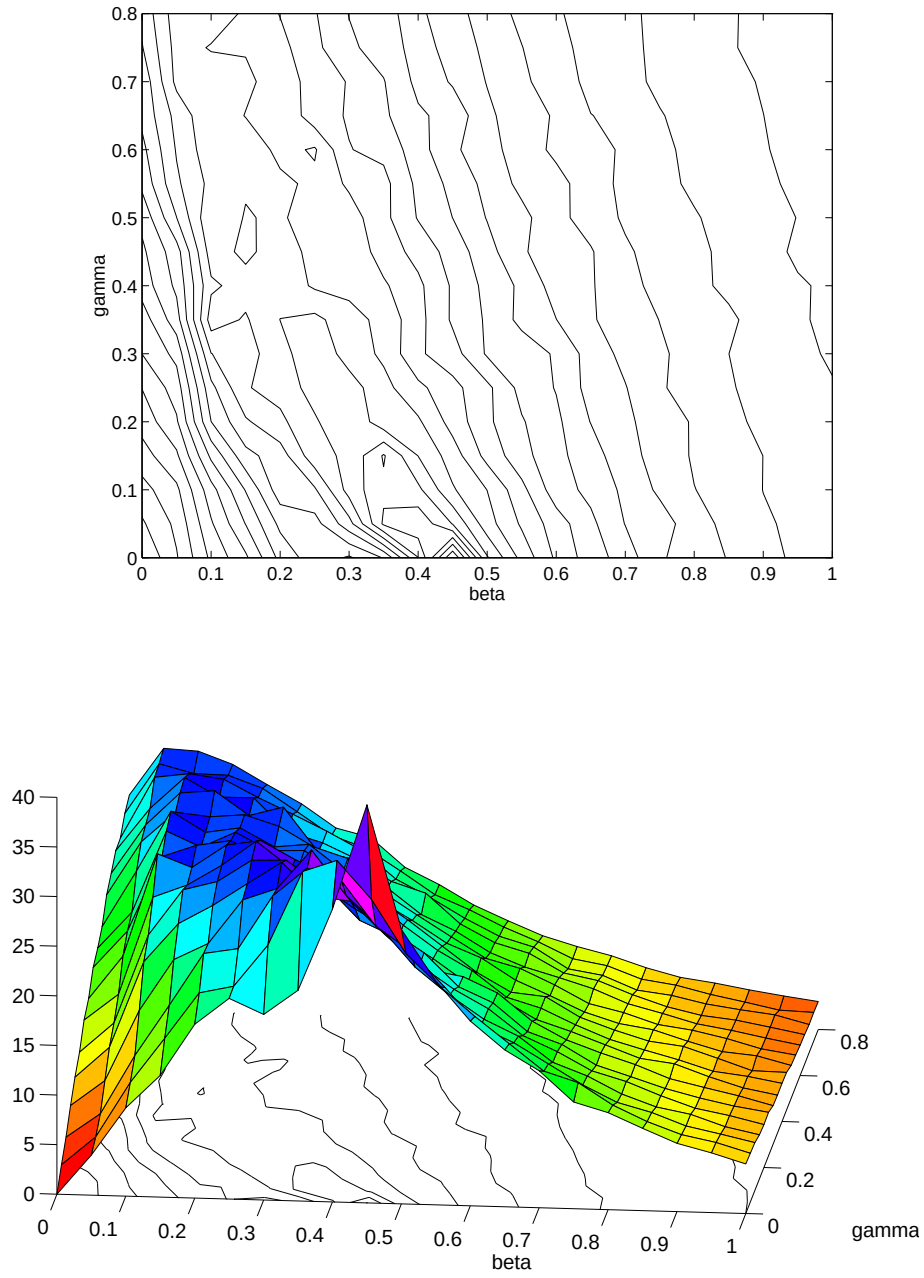

Figure 12: Plot of variance of energy function for hierarchical model

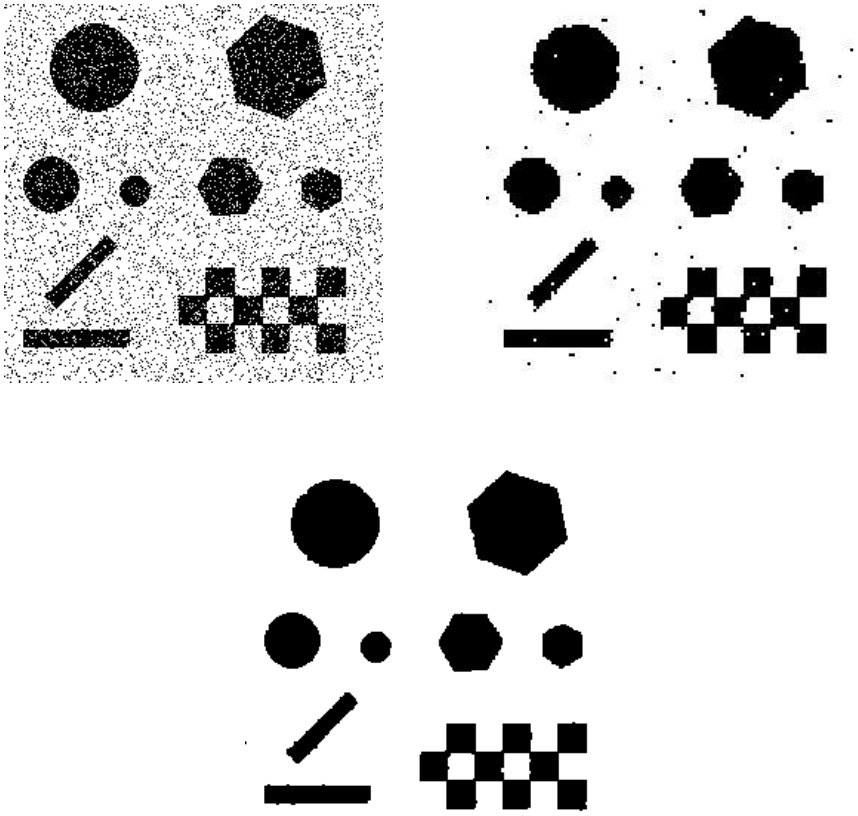

Figure 13: Test image with 10% channel noise (top left), segmentation with the hierarchical model with estimated parameters (top right) and segmentation with manually chosen parameters (bottom)

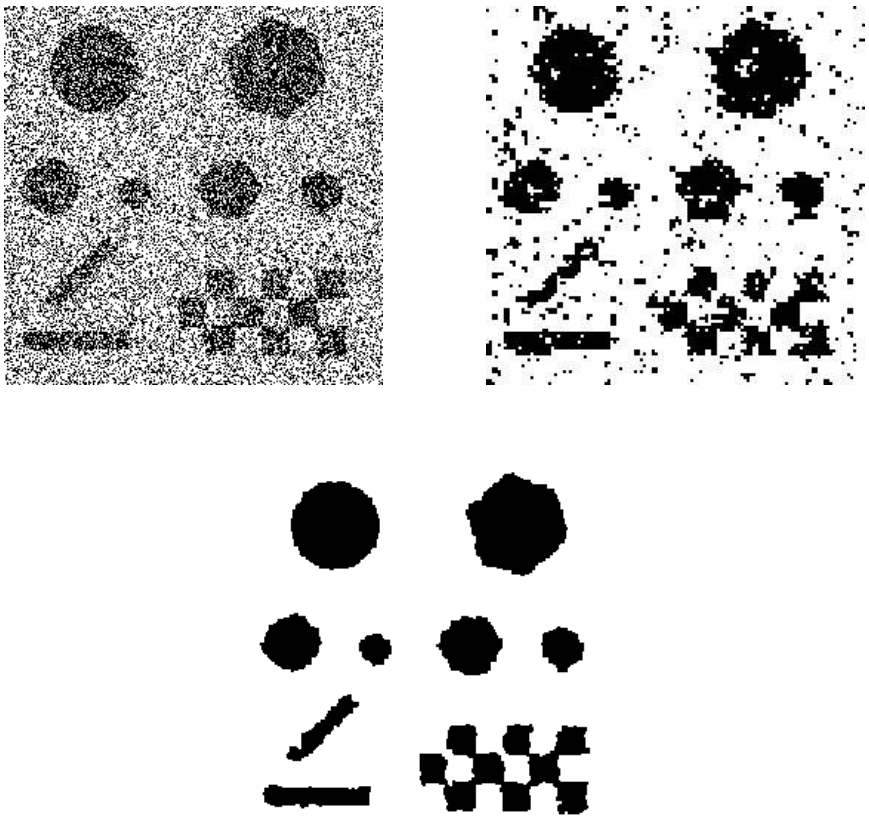

Figure 14: Test image with 30% channel noise (top left), segmentation with the hierarchical model with estimated parameters (top right) and segmentation with manually chosen parameters (bottom)

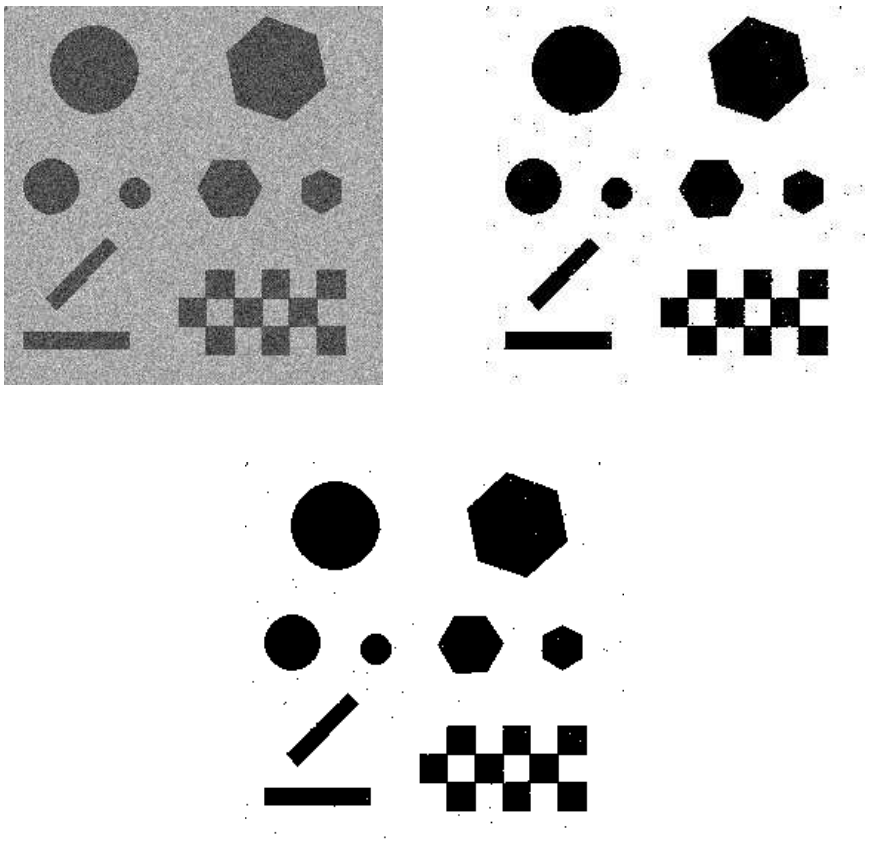

Figure 15: Gray level test image (top left), segmentation with the hierarchical model with estimated parameters (top right) and segmentation with manually chosen parameters (bottom)

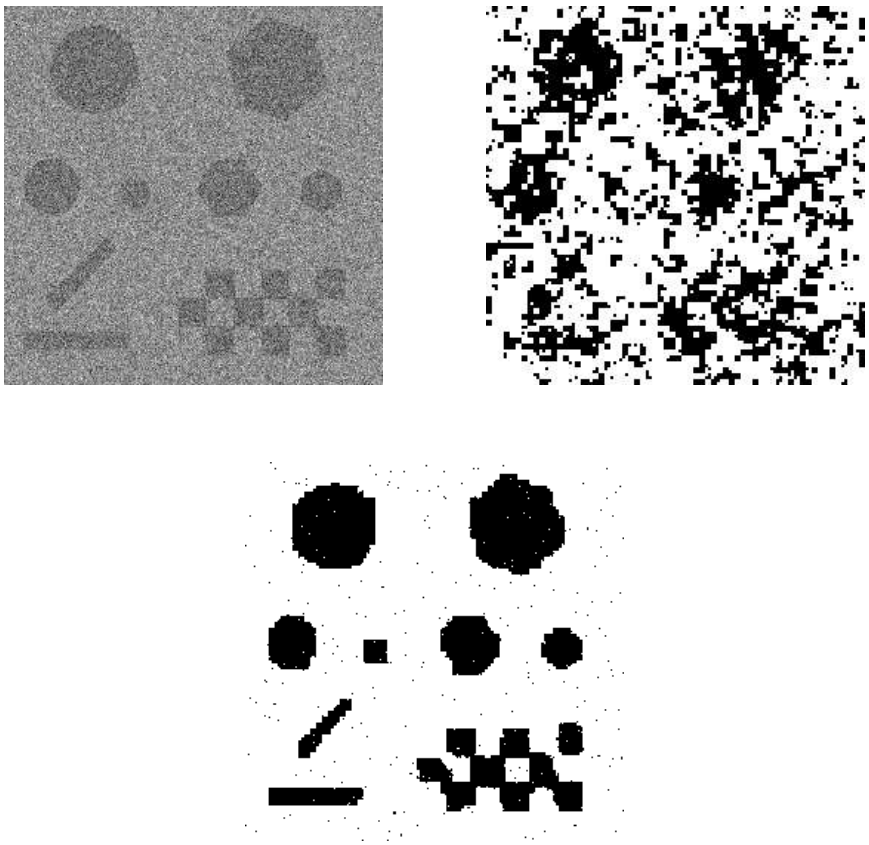

Figure 16: Gray level test image (top left), segmentation with the hierarchical model with estimated parameters (top right) and segmentation with manually chosen parameters (bottom)

for improved segmentation models (for example, models along the line of the chien model [4, 24]) is clear.

8 Conclusions

In this report we have studied some models used in the segmentation of satellite imagery. We have studied the problem of parameter estimation for these models, and shown that accurate parameter estimation is possible for these models, and indeed a much wider class of models than is usually thought. Accurate parameter estimation is a double edged sword, however – it often reveals that the characteristics of the images captured by the model do not correspond to the characteristics the user wishes to capture. This can often mean that ad-hoc parameter estimation schemes can give results which are closer to those desired by the user, but often accompanied by over-smoothing and loss of detail. The presence of a phase transition may also mean that care must be taken when generating samples from the models.

References

[1] J. Besag. On the statistical analysis of dirty pictures. *J. Royal Statistical Soc. B*, 48(3):259–302, 1986.

[2] B. Chalmond. Image restoration using an estimated Markov model. *Signal Processing*, 15:115–129, 1988.

[3] H. Derin and H. Elliott. Modelling and segmentation of noisy and textured images using Gibbs random fields. *IEEE Trans PAMI*, 9:39–55, 1987.

[4] X. Descombes, J-F. Mangin, E. Pechersky, and M. Sigelle. Fine structures preserving Markov model for image processing. In *Proceedings of SCIA-95*, pages 349–356, 1995.

[5] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans PAMI*, 6(6):721–741, November 1984.

[6] H-O Georgii. *Gibbs Measures and Phase Transitions.* de Gruyter, Berlin, 1988.

[7] C.J. Geyer and E.A. Thompson. Constrained Monte Carlo maximum likelihood for dependent data (with discussion). *JRSS-B*, 54:657–699, 1992.

[8] J. Goutsias. Markov random fields: Interacting particle systems for image modelling and analysis. Technical report, Department of Electrical and Computer Engineering, Image Analysis and Communications Laboratory, The Johns Hopkins University, Baltimore, MD, USA, 1996.

[9] W.K. Hastings. Monte Carlo sampling methods using Markov Chains and their applications. *Biometrika*, 57:97–109, 1970.

[10] D. Higdon. *Spatial Applications of Markov Chain Monte Carlo for Bayesian Inference.* PhD thesis, University of Washington, 1994.

[11] M. Hurn and C. Jennison. Multiple-site updates in maximum a-posteriori and marginal posterior modes image estimation. Technical Report 93:03, University of Bath, April 1993.

[12] Z. Kato. *Modélisations Markoviennes multirésolutions en vision par ordinateur. Appliction à la segmentation d'images SPOT.* PhD thesis, L'université de Nice-Sophia Antipolis, 1994.

[13] Z. Kato, M. Berthod, and J. Zerubia. A hierarchical Markov Random Field model and multitemperature annealing for parallel image classification. *Grapical Models and Image Processing*, 58:18–37, 1996.

[14] Z. Kato, J. Zerubia, and M. Berthod. Bayesian image classification using Markov random fields. In G. Demoments, editor, *Maximum entropy and Bayesian Methods*, pages 375–382. Kluwer, 1993.

[15] J. Marroquin, S. Mitter, and T. Poggio. Probabilistic solution of ill-posed problems in computational vision. *Journal of the American Statistical Association*, 82(397):76–89, March 1987.

[16] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, 21:1087–1092, 1953.

[17] R. Neal. Probabilistic inference using Markov Chain Monte Carlo methods. Technical Report CRG-TR-93-1, Department of Computer Science, University of Toronto, 1993.

[18] P. Perez and F. Heitz. Multiscale markov random fields and constrained relaxation in low level image analysis. In *Proceedings of ICASSP*. San Francisco, 1992.

[19] J.G. Propp and B.M. Wilson. Exact sampling with coupled markov chains and applications to statistical mechanics. Technical report, Massachusetts Inst. of Technology, 1995.

[20] R.D. Rosenkrantz, editor. *E. T. Jaynes: Papers on probability, statistics and statistical physics.* Kluwer, 1989.

[21] J.J.K. O Ruanaidh and W.J. Fitzgerald. *Numerical Bayesian Methods Applied to Signal Processing.* Springer-Verlag, Berlin, 1996.

[22] R.H. Swendsen and J-S. Wang. Nonuniversal critical dynamics in Monte Carlo simulations. *Physical Review Letters*, 58:86–88, 1987.

[23] L. Tierney. Markov chains for exploring posterior distributions. Technical Report 560, School of Statistics, University of Minnesota, 1991.

[24] H. Tjelmeland and J. Besag. Markov Random Fields with higher order interactions, 1996. Submitted to J. American Statistical Association.

Unité de recherche INRIA Lorraine, Technopôle de Nancy-Brabois, Campus scientifique,
615 rue du Jardin Botanique, BP 101, 54600 VILLERS LÈS NANCY
Unité de recherche INRIA Rennes, Irisa, Campus universitaire de Beaulieu, 35042 RENNES Cedex
Unité de recherche INRIA Rhône-Alpes, 655, avenue de l'Europe, 38330 MONTBONNOT ST MARTIN
Unité de recherche INRIA Rocquencourt, Domaine de Voluceau, Rocquencourt, BP 105, 78153 LE CHESNAY Cedex
Unité de recherche INRIA Sophia-Antipolis, 2004 route des Lucioles, BP 93, 06902 SOPHIA-ANTIPOLIS Cedex

Éditeur
INRIA, Domaine de Voluceau, Rocquencourt, BP 105, 78153 LE CHESNAY Cedex (France)
ISSN 0249-6399