J. R. Statist. Soc. B (1986)
48, *No.* 3, *pp.* 259–302

On the Statistical Analysis of Dirty Pictures

By JULIAN BESAG

University of Durham, U.K.

[*Read before the* Royal Statistical Society, *at a meeting organized by the* Research Section *on Wednesday, May 7th, 1986,* Professor A. F. M. Smith *in the Chair*]

SUMMARY

A continuous two-dimensional region is partitioned into a fine rectangular array of sites or "pixels", each pixel having a particular "colour" belonging to a prescribed finite set. The true colouring of the region is unknown but, associated with each pixel, there is a possibly multivariate record which conveys imperfect information about its colour according to a known statistical model. The aim is to reconstruct the true scene, with the additional knowledge that pixels close together tend to have the same or similar colours. In this paper, it is assumed that the *local* characteristics of the true scene can be represented by a non-degenerate Markov random field. Such information can be combined with the records by Bayes' theorem and the true scene can be estimated according to standard criteria. However, the computational burden is enormous and the reconstruction may reflect undesirable large-scale properties of the random field. Thus, a simple, iterative method of reconstruction is proposed, which does not depend on these large-scale characteristics. The method is illustrated by computer simulations in which the original scene is *not* directly related to the assumed random field. Some complications, including parameter estimation, are discussed. Potential applications are mentioned briefly.

Keywords: IMAGE PROCESSING; PATTERN RECOGNITION; SEGMENTATION; CLASSIFICATION; IMAGE RESTORATION; REMOTE SENSING; MARKOV RANDOM FIELDS; PAIRWISE INTERACTIONS; MAXIMUM A POSTERIORI ESTIMATION; SIMULATED ANNEALING; GIBBS SAMPLER; ITERATED CONDITIONAL MODES; GREY-LEVEL SCENES; AUTO-NORMAL MODELS; PSEUDO-LIKELIHOOD ESTIMATION.

1. INTRODUCTION

There has been much recent interest in the following type of problem. A continuous two-dimensional region S is partitioned into a fine rectangular array of sites or picture elements ("pixels"), each having a particular colour, lying in a prescribed set. The colours may be unordered, in which case they are usually tokens for other attributes associated with S, such as crop types in satellite images, or may be ordered, in which case they are usually grey levels and represent the value per pixel of some underlying variable, such as intensity. In either case, there is supposed to be a true but unknown colouring of the pixels in S and the aim is to reconstruct this scene from two imperfect sources of information.

The first of these is that, associated with each pixel, there is a possibly multivariate record which provides data on the colour there. It is assumed that, for any particular scene, the records follow a known statistical distribution. The second source is not directly quantitative but states that pixels close together tend to have the same colour or very similar colours, depending on whether colours are unordered or ordered. In this paper, we seek to quantify this second source probabilistically, by means of a non-degenerate Markov random field which crudely represents the *local* characteristic of the underlying scene. In principle, such an approach enables the two sources to be combined by Bayes' theorem and the true scene to be estimated according to standard criteria. For discrete colours, with which this paper is almost wholly concerned, the obvious choices are (i) the colouring that has overall maximum probability, given the

Present address: Department of Mathematical Sciences, University of Durham, Durham DH1 3LE, England.

 0035–9246/86/48259 $2.00

records, and (ii) that in which the colour at each individual pixel has maximum probability, given the records. In a Bayesian framework, (i) corresponds to maximum *a posteriori* estimation, whereas (ii) maximizes the posterior marginal probability at each pixel.

However, there are two reservations concerning (i) and (ii). The first is purely computational: scenes typically contain perhaps $2^{16} = 256 \times 256$ pixels, so that generally the optimization problems are enormous. Exceptions occur, by design, in the subclass of Markov random fields known as Markov mesh models (Abend *et al*, 1965) but such specifications are unnatural in a spatial context and can be very restrictive. Our second reservation is that, in choosing a non-degenerate field to describe the local properties of the scene, the induced large-scale characteristics of the model are somewhat undesirable. Thus, even relatively simple Markov random fields, such as the two-colour Ising model, in statistical physics, exhibit positive correlations over arbitrarily large distances when adjacent pixels are very likely to be the same colour; indeed, on the infinite lattice, there is a strong tendency to form *infinite* single-colour patches. Note that Monte Carlo simulations of Markov random fields are sometimes conditioned to have a *prescribed* number of pixels of each colour and hence the resulting patterns may be wholly unrepresentative of the corresponding *marginal* distributions.

The above reservations suggest that it may be useful to devise a method of reconstruction, which is still based on probabilistic considerations but whose computation is relatively trivial and which is unaffected by the large-scale characteristics of the chosen Markov random field. This is the main intention in the present paper.

In Section 2, we describe our basic assumptions and review the concept of a locally dependent Markov random field. We outline Geman and Geman's (1984) ingenious but computationally demanding solutions to (i) and (ii), through simulated annealing (Kirkpatrick *et al*, 1983) and the Gibbs sampler. We also discuss the Markov mesh family and indicate how particular models, especially that of Pickard (1977, 1980), have been used in tackling (ii) directly. Finally, we develop the simple iterative method of reconstruction, first proposed in Besag (1983) as an approximation to (i) but now considered in its own right. In Section 3, we illustrate this method on two artificial scenes which have been corrupted by white noise. The reconstructions involve the most basic choice of Markov random field, though this bears no direct relationship to the original scenes. Section 4 addresses wider classes of Markov random fields and suggests how they may be used to characterize different types of scenes. Section 5 examines some ways in which the basic assumptions can be relaxed and includes two further examples. Concluding remarks are in Section 6.

There are very many other methods of restoring corrupted images. The simplest, known as the maximum likelihood classifier, is of course to choose each pixel colour merely in accord with its own particular record or records. The resultant scene may be post-processed, using a local smoother such as simple majority vote, though this has the clear disadvantage of losing track of the records themselves. The same is true of the usual "relaxation labelling" techniques of, for example, Rosenfeld *et al* (1976) and Hummel and Zucker (1983); Kay and Titterington (1986) place these in a statistical context. In the signal processing literature, "regularization procedures" are used and specifically acknowledge the blurring associated with the "point spread function": a useful overview is given by Titterington (1985). Here, maximum entropy methods have been strongly promoted: see, for example, Gull and Daniel (1978), Burch *et al* (1983), Skilling (1984, 1986) and, for a statistician's viewpoint, Titterington (1984a). In fact, until very recently, there has been little interest among statisticians in image processing problems, with the notable exception of Paul Switzer, whose pioneering work, particularly in the context of satellite data (see, for example, Switzer, 1980, 1983), has done much to awaken others, including the present author. Finally, some further references, not mentioned elsewhere in the paper, include, among a vast literature, Fu and Yu (1980), Hand (1981), Devijver and Kittler (1982), Hansen and Elliott (1982), Rosenfeld and Kac (1982), Kittler (1983), Yu and Fu (1983), Hjort and Mohn (1984) and Saebo *et al* (1985).

Image processing is required in a very wide range of practical problems. Examples include:

various types of satellite data; images obtained from aircraft (e.g. synthetic aperture radar); ultra-sound; thermal images; nuclear medicine (e.g. photon emission tomography and scans obtained by nuclear magnetic resonance or gamma camera); computer vision (e.g. automatic object recognition); electron micrographs; and astronomy, where it is now common to collect data on a two dimensional array of detectors.

2. METHODOLOGY

2.1. *Notation and Assumptions*

We shall be dealing with rectangular arrays of pixels but it is notationally convenient, for the moment, merely to suppose that the (two-dimensional) region S has been partitioned into n pixels, labelled in some manner by the integers $i = 1, 2, \ldots, n$. Each pixel can take one of c colours, labelled $1, 2, \ldots, c$, with c finite. We assume there are no deterministic exclusions, so that the minimal sample space is $\Omega = \{1, 2, \ldots, c\}^n$: in Section 4, we briefly describe an exception and also the case of continuous "colours", which we then refer to as "intensities". An arbitrary colouring of S will be denoted by $x = (x_1, x_2, \ldots, x_n)$, where x_i is the corresponding colour of pixel i. We write x^* for the true but unknown scene and interpret this as a particular realization of a random vector $X = (X_1, X_2, \ldots, X_n)$, where X_i assigns colour to pixel i. We let y_i denote the observed record at i and y the corresponding vector, interpreted as a realization of a random vector, $Y = (Y_1, Y_2, \ldots, Y_n)$; recall that Y_i may itself have several components. For definiteness, we take Y_i to be continuous but the discrete case, perhaps where Y_i is itself a colour, is entirely analogous. We use $P(\cdot)$ and, on occasion, $P_T(\cdot)$ to denote probabilities of named events.

In developing a methodology, we shall formally make the following two assumptions, each of which we briefly discuss.

Assumption 1. Given any particular scene x, the random variables $Y_1, Y_2, \ldots, Y_n$ are conditionally independent and each Y_i has the same known conditional density function $f(y_i \mid x_i)$, dependent only on x_i. Thus, the conditional density of the observed records y, given x, is simply

$$l(y \mid x) = \prod_{i=1}^{n} f(y_i \mid x_i). \tag{1}$$

The assumption that f is known is usual and presupposes knowledge of the sensing mechanism and/or the existence of training data from known scenes. However, in Section 5, we shall consider parameter estimation *in situ*. Two modifications to Assumption 1 will be described in Section 5. First, as occurs with satellite data, there may be overlap between records, according to a known point-spread function, in that y_i may contain information not only from pixel i but also from adjacent pixels. The conditioning set in f must then be expanded to include the x_js at these pixels: we defer this primarily as a matter of notational convenience. Second, the assumption of conditional independence is not always valid: for example, the reflectances from adjacent pixels of wheat may be noticeably more alike than those from pixels further apart.

Assumption 2. The true colouring x^* is a realization of a locally dependent Markov random field (M.r.f.) with specified distribution $\{p(x)\}$.

We give a short description of locally dependent M.r.f.s below and discuss some specific examples in Sections 3 and 4. For practical purposes, the advance specification of a particular p may seem unrealistic but all we shall eventually require (Section 2.5) is that the distribution and the true scene are reasonably consistent as regards their *local* rather than their *global* characteristics. Incidentally, we might allow p to contain one or two unknown parameters and, in Section 5, we describe how these can be estimated during reconstruction.

2.2 *Locally Dependent Markov Random Fields*

Let $\{p(x)\}$ be a probability distribution which assigns colourings to S. Denote by x_A a colouring of the subset A of S and, in particular, by $x_{S\backslash i}$ a colouring of all pixels other than pixel i; of course, $x_S = x$ and $x_{\{i\}} = x_i$. Consider the conditional probability $P(x_i \mid x_{S\backslash i})$ of colour x_i occurring at pixel i, given the colouring $x_{S\backslash i}$ elsewhere; note that $S\backslash i$ is the only natural conditioning set for spatial distributions, as distinct from time series where time provides an obvious ordering. Viewed through its conditional distribution at each pixel, $\{p(x)\}$ is termed a *Markov random field.*

We shall be concerned only with fields whose conditional distributions are *locally dependent*; that is, depend only on the colours of pixels in the immediate vicinity of pixel i. Thus, suppose that, for every x,

$$P(x_i \mid x_{S\backslash i}) \equiv p_i(x_i \mid x_{\partial i}), \tag{2}$$

where p_i is specific to pixel i and ∂i is a subset of $S\backslash i$. Then the members of the set ∂i are termed the *neighbours* of pixel i. In practice, we approach the problem from the other end, by first naming the neighbours ∂i of each pixel i and then selecting $\{p(x)\}$ from among the corresponding class of probability distributions. Note that, when each pixel has only a few neighbours, this class is highly restricted by unobvious consistency conditions, identified by the Hammersley-Clifford theorem; see, for example, Besag (1974). Among the consequences of the theorem, it is necessary to preserve symmetry in naming neighbours: that is, if j is a neighbour of i then i must be a neighbour of j.

For a square array of pixels, there is an obvious hierarchy of virtually translation invariant neighbourhoods. Thus, the simplest departure from independence is a *first-order* M.r.f., in which the neighbours of each pixel comprise its available N, S, E and W adjacencies; boundary pixels have three rather than four neighbours, except at the corners where they have only two. The assumptions at the boundary are made for convenience. For a *second-order* field, the available pixels which are diagonally adjacent to pixel i are additionally included in ∂i; thus, each interior pixel has eight neighbours. The hierarchy can clearly be continued but, although the main methods in the paper apply to more general M.r.f.s, we shall only consider second-order fields in actual reconstructions; we view a first-order assumption to be unrealistic for most practical purposes. It should be stressed that we make no pretence of causal modelling in the specification (2). Equally, in a Bayesian framework, $\{p(x)\}$ is to be viewed merely as our prior distribution for the true scene x^*.

For further details of Markov random fields, see, for example, Besag (1974), and for a comprehensive mathematical treatment, Preston (1974) and Kinderman and Snell (1980).

2.3. *Maximum Probability Estimation*

In Sections 2.3 to 2.5, we consider some connected probabilistic methods of estimating the true scene x^*. In the first of these, the estimate $\hat{x}$ is chosen to have maximum probability, given the vector of records y. Thus, by Bayes' theorem, $\hat{x}$ maximizes

$$P(x \mid y) \propto l(y \mid x)p(x), \tag{3}$$

with respect to x. In a Bayesian framework, $\hat{x}$ is the maximum *a posteriori* (m.a.p.) estimate of x^*, being the mode of its posterior distribution; in the context of decision theory, $\hat{x}$ corresponds to a zero-one loss function, according to whether the reconstruction is perfect or imperfect, and is rather less appealing. The computational problem is clearly enormous and we therefore briefly describe how Geman and Geman (1984) tackle it, through the medium of *simulated annealing* (Kirkpatrick *et al.*, 1983).

Thus, consider a conceptual system in which any colouring x in Ω, for given records y, has probability

$$P_T(x \mid y) \propto \{l(y \mid x)p(x)\}^{1/T}, \tag{4}$$

where $T > 0$ is a parameter at our disposal and, in physical terms, represents the "absolute temperature" of the system. Note that, in the limit as $T \to \infty$, (4) is a uniform distribution on Ω; $T = 1$ corresponds to the distribution (3); and, in the limit as $T \to 0$, (4) is concentrated on the maximum probability estimate $\hat{x}$ of x^*, or, if multiple global maxima exist, is uniformly distributed over the corresponding $\hat{x}$s.

Unfortunately, it is not possible to simulate discrete M.r.f.s directly, except in very special cases (cf. Section 2.4). However, as shown below, it is easy to construct the transition matrix Q_T of a discrete time Markov chain, with state space Ω and limit distribution (4). Simulated annealing proceeds by running an associated time *inhomogeneous* Markov chain with transition matrices Q_T, where T is progressively decreased according to a prescribed "schedule" to a value close to zero. The final colouring is taken to maximize (3). The stochastic nature of the procedure enables escapes from any local maxima to occur but, to ensure such escapes, the chain must be run for a very long time and T must eventually be decreased extremely slowly. For proof of convergence and some practical recommendations concerning the "temperature" schedule, see Geman and Geman (1984).

There are various related methods of constructing a manageable Q_T (Hastings, 1970). Geman and Geman (1984) adopt the simplest, which they term the "Gibbs sampler". In the basic version, each pixel is addressed in turn in some prescribed order. When at pixel i, any colour x_i is chosen there with probability

$$P_T(x_i \mid y, x_{S\setminus i}) \propto \{f(y_i \mid x_i) p_i(x_i \mid x_{\partial i})\}^{1/T},$$

where $x_{S\setminus i}$ is the current colouring elsewhere; that is, x_i is chosen from its conditional distribution in (4). It is easily shown that the corresponding transition matrix for pixel i maintains (4), so Q_T can be taken as the product of these matrices over a complete cycle. Incidentally, time reversibility, a common ingredient in this type of problem (see, for example, Besag, 1977a), is present at individual stages but not over complete cycles, though Peter Green has pointed out that it returns if Q_T is taken over a pair of cycles, the second of which visits pixels in reverse order.

In fact, to ensure convergence of the Gibbs sampler and of simulated annealing, Geman and Geman (1984) make only weak demands on the manner in which pixels are visited. As regards the former (i.e. T fixed), any asynchronous method of updating, whether deterministic or stochastic, is acceptable, provided each pixel is visited infinitely often; and the procedure may even be synchronous to the extent that no two pixels which are neighbours should be simultaneously updated. As regards the latter, it is assumed that, at any stage, each pixel will be visited within an arbitrary but fixed number of updates. Note that partially synchronous updating relates to the division of pixels into "coding sets" (Besag, 1974) and is computationally important in the context of parallel processors. It should perhaps be stressed that, despite its capacity to reach (though not detect!) the global maximum of (3), simulated annealing is extremely simple to implement for locally dependent $\{p(x)\}$; when addressing pixel i, only y_i and the current colours of neighbours are required. On the other hand, it is therefore not surprising that an immense amount of computation is required for arrays of quite modest dimensions, let alone those generally encountered in practice.

2.4. *Classification by Maximum Marginal Probabilities*

In some circumstances, perhaps when constructing a crop inventory from satellite data for example, reconstruction of the scene itself may be of secondary importance. A more pertinent requirement may be to maximize the expected proportion of correctly classified pixels; that is, to estimate x_i^* for each i, by $\hat{x}_i$ which maximizes

$$P(x_i \mid y) \propto \sum_{x_{S\setminus i}} l(y \mid x) p(x),$$

the marginal (posterior) probability of x_i at i, given the records y. Note that $P(x_i \mid y)$ depends on all the records for (almost) any $\{p(x)\}$ and is extremely unwieldy. A more modest proposal is to

choose $\hat{x}_i$ to maximize $P(x_i \mid y_{\lambda i})$, where λi contains pixel i and prescribed associates; for example, λi might be a 3×3 block centred on pixel i. Even here, analytical progress is barred, in general, because $P(x_{\lambda i})$ is unavailable in closed form.

The exceptions to the general rule fall within the *Markov mesh* family (Abend *et al*, 1965; Kanal, 1980), a class of models designed specifically with this aim. Suppose the pixels of a rectangular array are now labelled (i, j), where i and j refer to row and column, respectively. Any pixel (i', j') is called a *predecessor* of (i, j) if $i' < i$, or $i' = i$ and $j' < j$. A Markov mesh model is defined by specifying a conditional distribution for each X_{ij}, given the colours of its predecessors: this distribution is deemed to depend only on a small set R_{ij} of predecessors. The unilateral nature of the models implies that the unobvious consistency conditions of Section 2.2 do not arise and that direct simulation can be carried out in a raster scan along successive rows.

In the simplest examples of interest, $R_{ij} = \{(i-1, j), (i, j-1)\}$, for any pixel beyond the first row and column; see Bartlett and Besag (1969), Galbraith and Walley (1976), and, for applications to image processing, Derin *et al* (1984) and Devijver (1985). For any Markov mesh model, the conditional distribution of X_{ij}, given all other pixel colours, can be easily found and, in this particular case, each interior pixel (i, j) has six neighbours: those immediately adjacent to it are augmented by $(i-1, j+1)$ and $(i+1, j-1)$. Bartlett and Besag (1969) and Besag (1974) discuss how the addition of $(i-1, j+1)$ to R_{ij} above can be used to partially eliminate the undesirable spatial asymmetry but this device will fail when substantial local dependence is present. The asymmetry pervades the family as a whole and it is not clear how to incorporate particular spatial attributes: indeed, Markov mesh models are perhaps most appropriately thought of as two-dimensional time series.

The one exception to the above asymmetry is provided by the Pickard (1977, 1980) model, adopted in image processing, implicitly by Kittler and Föglein (1984a) and explicitly by Haslett (1985); see also Derin *et al* (1984). The model has the curious property that, given the colour of any pixel, the colours of immediately adjacent pixels are conditionally independent. It is a very special example of a Markov mesh model with $R_{ij} = \{(i-1, j), (i, j-1), (i-1, j-1)\}$ and of a second-order M.r.f. Since, for example, the conditional distribution of X_{ij}, given the colouring of all columns $j' < j$, depends only on $x_{i,j-1}$, it follows that the Pickard model lays emphasis inappropriately on the orientation of the pixel array with respect to the scene. In mitigation, computation is much simplified, though it can still be expensive in time and storage. Thus, reverting to our general pixel notation, suppose λi is chosen to include pixel i and its first-order neighbours. Then,

$$P(x_{\lambda i}) = P(x_i) \prod_{j \in \lambda i \setminus i} P(x_j \mid x_i).$$

Haslett (1985) extends λi to include all pixels in the same row and column as pixel i and gives an approximation to include second-order neighbours.

Finally, in this discussion, we note the link with Section 2.3. As Geman and Geman (1984) point out, any property of the (posterior) distribution $P(x \mid y)$ can be simulated by running the Gibbs sampler at "temperature" $T = 1$. Thus, if $\hat{x}_i$ maximizes $P(x_i \mid y)$, then it is the most frequently occurring colour at pixel i in an infinite realization of the Markov chain with transition matrix Q_1 of Section 2.3. The $\hat{x}_i$s can therefore be simultaneously estimated from a single finite realization of the chain. It is not yet clear how long the realization needs to be, particularly for estimation near colour boundaries, but the amount of computation required is generally prohibitive for routine purposes.

2.5. *Estimation by Iterated Conditional Modes (ICM)*

We have seen that, if there is to be a flexible choice of $\{p(x)\}$, the methods of Sections 2.3 and 2.4 make enormous computational demands. There is an additional problem: although we can easily specify first- and particularly second-order M.r.f.s $\{p(x)\}$ on S, which have appealing

local characteristics (Sections 3 and 4), these same fields generally have undesirable large-scale properties. In particular, given the level of local dependence we have in mind, realizations of $\{p(x)\}$ on S will usually consist almost entirely of a single colour. Of course, this will not generally be true of $P(x \mid y)$ because of the influence of the records; nevertheless, one might prefer a method of reconstruction which is not only computationally undemanding but also ignores the large scale deficiencies of $\{p(x)\}$. We now describe a scheme which satisfies these conditions.

Suppose that $\hat{x}$ denotes a provisional estimate of the true scene x^* and that our aim is merely to update the current colour $\hat{x}_i$ at pixel i in the light of all available information. Then a plausible choice is the colour which has maximum conditional probability, given the records y and the current reconstruction $\hat{x}_{S\setminus i}$ elsewhere; that is, the new $\hat{x}_i$ maximizes $P(x_i \mid y, \hat{x}_{S\setminus i})$ with respect to x_i. It follows from Bayes' theorem and equations (1) and (2) that

$$P(x_i \mid y, \hat{x}_{S\setminus i}) \propto f(y_i \mid x_i) p_i(x_i \mid \hat{x}_{\partial i}), \tag{5}$$

so that implementation is trivial for *any* locally dependent M.r.f. $\{p(x)\}$. When applied to each pixel in turn, this procedure defines a single cycle of an iterative algorithm for estimating x^*.

As an initial $\hat{x}$, we shall normally adopt the conventional maximum likelihood classifier, which ignores geometrical considerations and merely chooses $\hat{x}_i$ to maximize $f(y_i \mid x_i)$ at each i separately. We then apply the algorithm for a fixed number of cycles or until convergence, to produce the final estimate of x^*: note that

$$P(x \mid y) = P(x_i \mid y, x_{S\setminus i}) P(x_{S\setminus i} \mid y),$$

so that $P(\hat{x} \mid y)$ never decreases at any stage and eventual convergence is assured. In practice, convergence, to what must therefore be a local maximum of $P(x \mid y)$, seems extremely rapid, with few if any changes occurring after about the sixth cycle. Indeed, it was as an approximation to maximum probability estimation that the algorithm was first proposed (Besag, 1983), although we no longer view it merely in that light. The algorithm was suggested independently by Kittler and Föglein (1984b), who applied it to Landsat data, as did Kiiveri and Campbell (1986). Note that its dependence only on the local characteristics of $\{p(x)\}$ is ensured by the rapid convergence. We label the method *ICM*, representing "iterated conditional modes".

The actual mechanics of updating may depend on computing environment. Thus, with a language such as Fortran, updating is most conveniently implemented as a raster scan. It is helpful to vary the raster from cycle to cycle, in order to reduce the small directional effects which may otherwise be produced. However, with a matrix language such as APL, it will be much faster and more convenient to modify the algorithm and use synchronous updating: that is, to update cycle by cycle. This eliminates spurious directional effects but convergence can no longer be guaranteed and small oscillations may occur. The partially synchronous scheme, in which coding sets of pixels are simultaneously updated, provides a useful compromise. Given the ideal of a genuinely parallel system, with a single processor dedicated to each pixel, the processors could be allowed to run at their own individual rates and restoration would be immediate.

Finally, as remarked by Peter Green, there is a somewhat incidental link with Section 2.3. Apart from the totally synchronous option, which is invalid for the Gibbs sampler, *ICM* is exactly equivalent to instantaneous freezing in simulated annealing!

2.6. *Some Modifications of ICM*

There are several modifications which can be made to the basic version of *ICM*. A simple variant is to work up to the chosen $\{p(x)\}$, using a sequence of weaker fields on previous cycles. This can have the advantage of not fixing pixel colours too early in the restoration, on the basis of the unreliable initial estimate, though it naturally requires a little more iteration. Examples are provided in Sections 3 and 5.

A more radical suggestion is to replace single pixel optimization in (5) by maximization over a (small) set of contiguous pixels; that is, for a particular sequence of sets B in S, choose as $\hat{x}_B$ that x_B which maximizes $P(x_B \mid y, \hat{x}_{\partial B})$, where ∂B denotes the neighbours of all pixels in B which do not themselves belong to B and $\hat{x}_{\partial B}$ denotes the current reconstruction there. Extremes of this are of course $B = \{i\}$, which is (5), and $B = S$, which is (3). We have in mind, perhaps, blocks B of four or five pixels, though even here exhaustive search can be time consuming; see Section 5 for further discussion and a two-colour example. Note that, with asynchronous updating, any of the foregoing procedures terminates at a local maximum of $P(x \mid y)$.

Finally, it is possible to incorporate some degree of parameter estimation in $\{p(x)\}$, though this may be both computationally expensive and, perhaps surprisingly, unnecessary in applications. We provide further details and examples in Section 5.

3. EXAMPLES OF *ICM*

This section contains two artificial examples. Each involves a known scene x^* on an array of square pixels: at each pixel i, a univariate record y_i is generated as an observation from a Gaussian distribution with specified mean $\mu(x_i^*)$ and variance κ. The records are then processed by *ICM* (Section 2.5), using a very simple form of $\{p(x)\}$, to produce an estimate $\hat{x}$ of x^*.

In Section 4, we describe a quite widely applicable class of M.r.f.s but here we adopt the simplest non-degenerate version, with c unordered colours and a second-order neighbourhood. This assigns conditional probability

$$p_i(k \mid \cdot) = e^{\beta u_i(k)} \Big/ \sum_{l=1}^{c} e^{\beta u_i(l)} \tag{6}$$

to colour k at pixel i, given the colouring elsewhere, where $u_i(l)$ denotes the number of neighbours of i having colour l, and β is a fixed parameter, which, when positive, encourages neighbours to be of like colour. We choose $\beta = 1.5$ in the reconstructions: this value seems to work well, also in other examples (two of which appeared in the original version of the paper), and is supported by informal calculations, though the precise value is not crucial and there are arguments for a somewhat smaller choice. No appreciable gain was noted in allowing a reduced value for diagonal adjacencies but this is probably because our true scenes are tied to the lattice structure rather than a genuine continuous region. In a Bayesian framework, (6) is an *exchangeable* prior and we might take $\beta = 1.5$ to represent prior ignorance about the composition of the true scene.

In the above setting, a single cycle of *ICM* requires the successive minimization of

$$\frac{1}{2\kappa}\{y_i - \mu(x_i)\}^2 - \beta \hat{u}_i(x_i), \tag{7}$$

with respect to x_i, for each pixel in turn, where $\hat{u}_i(x_i)$ is the current number of neighbours of i having colour x_i.

Note that in the same setting, maximum probability (m.a.p.) estimation requires the simultaneous minimization of

$$\frac{1}{2\kappa}\sum_{i=1}^{n}\{y_i - \mu(x_i)\}^2 - \beta v, \tag{8}$$

with respect to the x_is, where v is the number of neighbour pairs having like colours: this follows from the fact that $p(x) \propto \exp(\beta v)$, as in Section 4.1.1. One suspects that, in this case, a direct minimization of (8) is available, thus avoiding simulated annealing.

Both (7) and (8) are intuitively appealing: in each, the first term matches the fitted colour mean to the record and the second gives a prize for smoothness. Note that (8) is a penalized log-likelihood and that other interpretations are available; see, for example, Titterington (1985).

These observations extend to multivariate Gaussian records, for which the quadratic terms in (7) and (8) are replaced by the appropriate generalized quadratic forms. Finally, it is of some interest to examine the extremes of β in (7) and (8). Of course, $\beta = 0$ gives the maximum likelihood classifier, with which we initialize *ICM*: this merely chooses the colours with means closest to the corresponding records, or, in the multivariate case, applies linear discriminant analysis. However, as $\beta \to \infty$, the two methods are radically different. Thus, *ICM* provides a recursive majority-vote solution acting on the initial closest-mean reconstruction, with records used only to break ties; whereas (8) results in a single-colour solution, the colour being that most frequent in the closest-mean reconstruction. Incidentally, it is plausible that, in general, a rather weaker $\{p(x)\}$ should be used in maximum probability reconstructions than in *ICM*; in particular, a value of β substantially less than 1.5 would be appropriate in (8).

Example 1. The true scene in the first example is shown in Fig. 1a. There are only three "colours", on a 90 × 98 array, but the scene contains some interesting features and, in its overall structure, is *not* typical of any simple M.r.f. Records were generated from independent

Fig. 1a. True three-colour scene: 90 × 98.

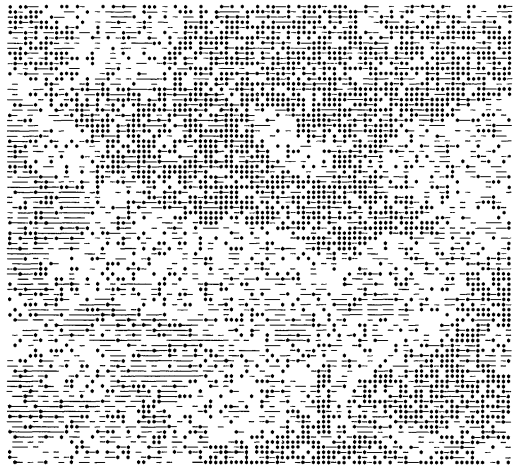

Fig. 1b. Maximum likelihood classifier: 40% error rate.

Fig. 1c. ICM reconstruction with $\beta = 1.5$: $< 5\%$ error rate.

circular Normal distributions, with means located at 0°, 120° and 240°, according to the true colour. With so few colours, this device aids interpretation by removing spurious edge effects and also enables κ to be specified to produce any desired mean level of misclassified pixels in the initial closest-mean reconstruction. Here this level was chosen as 40%; see Fig. 1b, from which it is difficult to discern the underlying true pattern.

Fig. 1c shows the result of applying six cycles of *ICM*, each with $\beta = 1.5$, and gives a reasonably faithful representation of the original scene, with less than 5% misclassification. The retention of the peninsula towards the bottom of the scene is quite encouraging. Incidentally, the program was written in APL and so it was natural to use synchronous updating.

Example 2. Here the true scene contains six "colours", on a 120×120 array; see Fig. 2a. It was originally hand-drawn and chosen to display a wide variety of features, including large and small patches, straight and curved boundaries between colours, sharp and rounded turns, and colours occurring with widely differing frequencies. A colour key, which is part of the pattern, is shown at the top right of Fig. 2a: thus, "minus" = 1, "cross" = 2, and so on. The records were generated from the colour labels by superimposing independent Gaussian noise, with standard deviation $\sqrt{\kappa} = 0.6$. It should be noted that the scene is perhaps a little dishonest, with a tendency for adjacent patches not to have adjacent colour labels: this clearly enhances subsequent restoration.

The processing was carried out on a standard 32K BBC Model B micro-computer. Fig. 2b shows the initial reconstruction, with colours whose labels are linearly closest to the records, and has a 34% error rate. The error rates after 1, 2, 4 and 6 cycles of *ICM*, with $\beta = 1.5$ throughout, were 3.6, 2.7, 2.0 and 1.7%, respectively. However, we also applied the first modification of *ICM* in Section 2.6, with eight complete cycles and β increasing by equal increments from 0.5 to 1.5 over the first six. This resulted in an improvement, with error rates 6.4, 2.7. 1.3, 1.0 and 0.9% after 1, 2, 4, 6 and 8 cycles, respectively: the program was run in BASIC, with records accessed from disk: reconstruction was carried out in Mode 2 by addressing the screen itself in a raster scan, so that very little storage, apart from screen memory (20K), was required.

We infer from the above and other examples that *ICM* can usefully build spatial considerations into the information directly available from the records. For example, it seems unlikely that reconstructions of the quality of Figs 1c and 2c could be obtained by direct visual or automatic smoothing of Figs 1b and 2b. Of course, some aspects of the examples are much simpler than would normally be met in practice: thus, in Sections 4 and 5, we consider some complicating factors.

4. MODELS FOR THE TRUE SCENE

This section is concerned with the specification of $\{p(x)\}$ for three different types of scenes: those with unordered colours, as in Section 3; grey-level scenes, either with smooth variation or with occasional abrupt changes in level; and those with continuous pixel intensities, either smoothly varying or with substantial disparities. Although we consider each at a rather basic level, even then, some of our comments are speculative. More intricate modelling, which takes account of special structures in the true scene, is a topic which requires much further research.

We shall assume that pixels have prescribed neighbours, satisfying the simple symmetry condition and chosen with proper regard to spatial contiguity: in fact, although not necessary, we shall have in mind a rectangular array of pixels with at least a second-order neighbourhood. We concentrate on *pairwise interactions*: $\{p(x)\}$ is called a pairwise interaction M.r.f. if, for any x in the sample space Ω,

$$p(x) \propto \exp\left\{\sum_{1 \leqslant i \leqslant n} G_i(x_i) + \sum\sum_{1 \leqslant i < j \leqslant n} G_{ij}(x_i, x_j)\right\}, \tag{9}$$

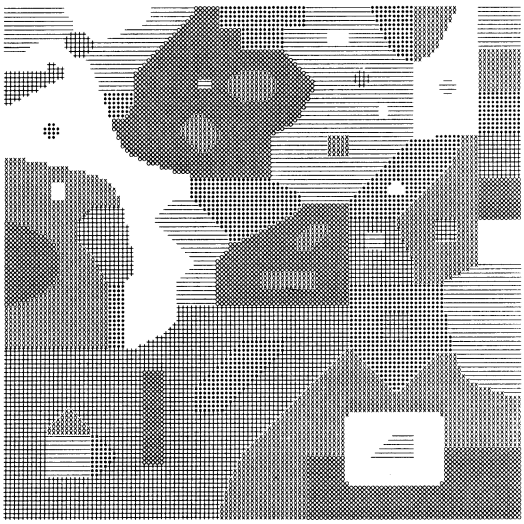

Fig. 2a. True six-colour scene: 120 × 120.

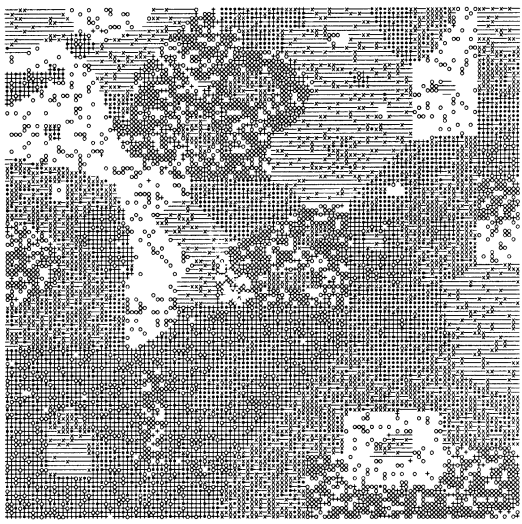

Fig. 2b. Maximum likelihood classifier: 34% error rate.

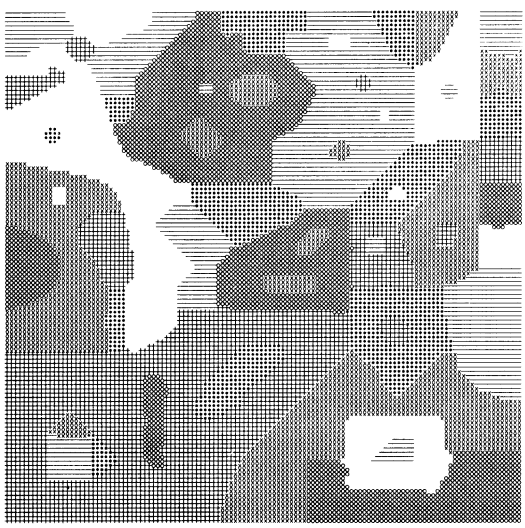

Fig. 2c. ICM reconstruction with $\beta \uparrow 1.5$: < 1% error rate.

where, to ensure the Markov property, $G_{ij} \equiv 0$ unless pixels i and j are neighbours but otherwise the G-functions are arbitrary (cf. Besag, 1974, with a trivial change of notation).

4.1. *Discrete colours*

As a special case of (9), we consider c-colour distributions with (Besag, 1976; Strauss, 1977)

$$p(x) \propto \exp\left(\sum_{1 \leqslant k \leqslant c} \alpha_k n_k - \sum\sum_{1 \leqslant k < l \leqslant c} \beta_{kl} n_{kl}\right), \tag{10}$$

where n_k denotes the number of pixels coloured k and n_{kl} the number of distinct neighbour pairs coloured (k, l); note that, for the moment, all neighbour pairs are treated equally. The αs

and βs are arbitrary parameters, though we shall be concerned only with $\beta_{kl} > 0$, which discourages colours k and l from appearing at neighbouring pixels. Further assumptions about the βs will be made below, according to the type of scene. Note there is a redundancy among the αs, since $n_1 + \ldots + n_c = n$, but we retain the symmetric formulation for neater exposition.

By considering any two realizations which differ only at pixel i, it follows that the conditional probability of colour k occurring there, given the colours of all other pixels is

$$p_i(k \mid \cdot) \propto \exp\{\alpha_k - \sum_{l \neq k} \beta_{kl} u_i(l)\}, \tag{11}$$

where $\beta_{kl} \equiv \beta_{lk}$ defines β_{kl} for $k > l$, and $u_i(l)$ denotes the number of neighbours of pixel i having colour l; incidentally, this confirms the Markov property of (10). The corresponding conditional probabilities for small sets of pixels, rather than a single pixel, can be found in a similar manner. However, marginal probabilities are generally intractable and the same is true even for the normalizing constant in (10); indeed, this has been the source of much anguish in statistical physics over many decades. Analytical results seem available only for the Ising model, which is the (infinite) rectangular lattice version of (10), with only two colours and first-order neighbours. The physicists' interest in such models centres on their capacity to produce positive correlations between the colours of pixels infinitely far apart, and hence infinite patches of a single colour. Fortunately, the normalizing constant is not required for maximum probability (m.a.p.) estimation or for *ICM*; moreover, we have designed the latter to be unresponsive to the large-scale characteristics of $\{p(x)\}$.

A simple extension of (10) is to allow differing interaction parameters β_{kl} to be associated with different classes of neighbour pairs. Thus, with square pixels and a second-order M.r.f., $\beta_{kl} n_{kl}$ in (10) can be replaced by $\beta'_{kl} n'_{kl} + \beta''_{kl} n''_{kl}$, where n'_{kl} and n''_{kl} are the respective numbers of first- and second-order neighbour pairs coloured (k, l); in this context, Geman and McClure (1985) downweight diagonal adjacencies by a factor of $\sqrt{2}$. A similar modification can be made for $N - S$ and $E - W$ adjacencies if pixels are rectangular rather than square. In the sequel, we assume that any such extensions would be incorporated when necessary: the corresponding amendments to (11) are immediate.

4.1.1. *Unordered colours*

A simple model for unordered colours is obtained by taking a common β in (10). Then (11) becomes

$$p_i(k \mid \cdot) \propto \exp\{\alpha_k + \beta u_i(k)\}, \tag{12}$$

since $u_i(1) + \ldots + u_i(c)$ is the total number of neighbours of pixel i. Note that (12) implies that the conditional *odds*, in favour of colour k, depend only on the number of like-coloured neighbours, not that this is true of the corresponding *probability*. A final simplification occurs if the colours are interchangeable, in which case the αs must also be equal and can be set to zero because of the inherent redundancy: this yields the distribution (6) used in Section 3.

4.1.2. *Excluded adjacencies*

A different situation arises if it is known that certain colours, say colours 1 and 2, cannot appear on neighbouring pixels in the true scene. This can be enforced in reconstructions by letting $\beta_{12} \to \infty$ in (10), by which we mean $\beta_{12} n_{12}$ is zero if $n_{12} = 0$ but is otherwise infinite. Such fields strictly contravene the positivity condition in the Hammersley-Clifford theorem and, in using *ICM*, synchronous updating should be avoided and the ostensibly infinite βs gradually increased cycle by cycle, particularly if the initial estimate is infeasible.

4.1.3. *Grey-level scenes*

If the colours have a natural ordering, usually ranging from black to white, we refer to them as grey levels. Image processing equipment typically has $2^8 = 256$ levels, in which case it may

be preferable to model the scene by a continuous $\{p(x)\}$ and then bin accordingly. Otherwise we can again appeal to (10), though it is convenient here, and necessary in discussing Cross and Jain (1983) below, to append a zero level (black, say) to the levels 1 to c. Then, regarding levels as "equally spaced", we might set $\beta_{kl} \equiv b_{|k-l|}$, where $b_1, b_2, \ldots, b_c$ is a specified sequence of non-decreasing constants. To encourage smooth variation, the sequence should be strictly increasing; whereas, if occasional abrupt changes in level are anticipated, the coefficients should quickly reach a maximum.

Our experience of *ICM* on such scenes is limited to a couple of artificial examples and several lung and liver scans produced by a gamma camera, with Poisson distributed records. The processing was carried out on the BBC Model B, with 8 or 16 levels and second-order neighbourhood. Although the results were satisfactory, it would seem desirable generally to adopt a larger neighbourhood to allow for curvature effects, through coefficients which depend also on the relative positions of neighbours. In any case, the medical examples really demand 256 grey levels and we do not yet have the computing equipment with which to process, display and experiment at this resolution.

A different class of distributions, with pixel outcomes 0, 1, ..., c, is provided by the auto-binomial family (Besag, 1974): $\{p(x)\}$ is auto-binomial if it satisfies (9) and has conditional distribution $p_i(x_i \mid x_{\partial i})$, at each pixel i, which is binomial with parameters c and $\theta_i(x_{\partial i})$, say. It is shown in Besag (1974) that, for $\{p(x)\}$ to be valid, the θs must satisfy

$$\theta_i(x_{\partial i})/\{1 - \theta_i(x_{\partial i})\} = \exp\left(\alpha_i + \sum_{j \neq i} \beta_{ij} x_j\right),$$

where $\beta_{ij} \equiv \beta_{ji}$ and $\beta_{ij} = 0$ unless pixels i and j are neighbours.

Cross and Jain (1983) provide a highly diverse collection of scenes from auto-logistic ($c = 1$) and auto-binomial models, using a variety of neighbourhoods and parameter values, and up to 32 grey levels. However, it should be noted that all their simulations were conditioned to have prescribed numbers of pixels at each level and many would have looked qualitatively different and less interesting had unconditional simulation been used. In fact, Cross and Jain's aim was to produce specific models whereby real textures, such as those of cork or paper, could be recognized automatically. Cross and Jain used the "coding" technique (Besag, 1974) to estimate neighbourhoods and parameter values and hence their models do indeed capture the *local* characteristics of the textures; see Section 5.1 for some further comments.

4.2. *Continuous Intensities*

4.2.1. *Auto-Normal models*

The simplest description of continuous intensities, perhaps after suitable transformation, is that $\{p(x)\}$ is a Gaussian M.r.f. with zero mean. Thus (Besag 1974, 1975), X_i has conditional density

$$p_i(x_i \mid x_{\partial i}) \propto \exp\left\{-\frac{1}{2\lambda_i}(x_i - \sum_{j \neq i} \beta_{ij} x_j)^2\right\}, \tag{13}$$

where (i) $\beta_{ij} = 0$ unless pixels i and j are neighbours, (ii) $\beta_{ij}\lambda_j = \beta_{ji}\lambda_i$ and (iii) an extra regularity condition below is satisfied. For second-order neighbourhood, the conditional variances λ_i would usually be taken as equal for all interior pixels but might be increased on the edges of the array and particularly at the corners, because less information is available there. It follows from (13) that

$$p(x) \propto \exp(-\tfrac{1}{2}x^{\mathrm{T}}Qx),$$

where $Q = \Lambda^{-1}B$, Λ is the $n \times n$ diagonal matrix with diagonal entries λ_i, and B is the $n \times n$ matrix with unit diagonal entries and off-diagonal (i, j) element $-\beta_{ij}$. Condition (ii) ensures

that the precision matrix Q is symmetric, while (iii) is a statement that Q must be positive definite.

It is of some interest to see how the above formulation interacts with Gaussian records. Thus, suppose that each pixel i produces an independent record y_i, generated by a Normal distribution with mean x_i^* and variance κ. Then, by various criteria, including m.a.p. estimation (Section 2.3), the estimate $\hat{x}$ of the true scene x^* is chosen to maximize the (posterior) expectation of X given y; that is, $\hat{x}$ minimizes

$$\kappa^{-1}(y - x)^{\mathrm{T}}(y - x) + x^{\mathrm{T}}Qx, \tag{14}$$

with respect to x (cf. least squares smoothing in Green, Jennison and Seheult, 1985), and hence

$$\hat{x} = (I + \kappa Q)^{-1}y, \tag{15}$$

where I is the $n \times n$ identity matrix. Furthermore, since (14) has a single minimum, it follows from Section 2.5 that *ICM* must converge to the same solution. Specifically, the updating formula at pixel i is

$$\hat{x}_i = (\lambda_i y_i + \kappa \sum_{j \neq i} \beta_{ij}\hat{x}_j)/(\lambda_i + \kappa), \tag{16}$$

a linear combination of the signal at i and the current estimates $\hat{x}_j$ at neighbouring pixels. Thus, *ICM* implements the matrix inversion in (15) by a sequence of local operations: convergence will therefore be rather slow (cf. discrete colours in Section 2.5). In the light of earlier discussion, the equivalence between m.a.p. estimation and *ICM* may seem embarrassing, save that no such Gaussian model can sustain positive correlations over infinite distances. Note that (16) highlights the fact that (13) is tailored for smooth variation in the true scene: non-linear versions of (16) are clearly of considerable practical interest.

4.2.2. *Other distributions*

The continuous analogue of the grey-level models in Section 4.1.3 is the family of pairwise interaction distributions (9) with $G_{ij}(x_i, x_j) \equiv -\phi_{ij}(x_i - x_j)$, where $\phi_{ij}(z)$ is even and non-decreasing in $|z|$. Of course, $\phi_{ij}(z) \equiv 0$ unless i and j are neighbours but might otherwise depend on the relative positions of pixels i and j. This family is adopted by Geman and McClure (1985), in the context of single photon emission tomography (*SPET*).

The auto-Normal formulation (13) does not quite fall into the above class, except in the intuitively appealing extreme case where $\Sigma\beta_{ij} = 1$ for each i, when $\phi_{ij}(z) = \frac{1}{2}\gamma_{ij}z^2$, with $\gamma_{ij} = \beta_{ij}/\lambda_i = \beta_{ji}/\lambda_j$. The row and column sums of Q are then zero, Q is only positive *semi*-definite and $\{p(x)\}$ is "just" improper. Nevertheless (15) and (16) are still valid.

The impropriety extends to any member of the family with unbounded support for each x_i. Again this need not be of practical consequence and, in any case, a restricted range can be imposed. Thus, to encourage sudden jumps in intensity, Geman and McClure adopt the form

$$\phi_{ij}(z) = \beta_{ij}/\{1 + (z/\delta)^2\},$$

in a range of intensities dictated by the image processing equipment, where δ is a tuning constant and β_{ij} depends on the relative positions of pixels i and j. There is much scope for experimentation with other choices, including more overtly resistant distributions and those with a simple conditional probability interpretation. As regards reconstruction, modal criteria, especially in a global sense, may be suspect. In fact, the impressive results in Geman and McClure (1985) are obtained by gradient descent* of the objective function, starting from the maximum likelihood, non-spatial solution, and presumably do not reach the global "optimum".

4.3. *Special features*

If $\{p(x)\}$ is to support special features, such as thin lines, then it is inevitable that pairwise

* It transpires that Geman and McClure used ICM and not gradient descent.

interaction models will not suffice. For some suggestions, see Grenander (1983) and Geman and Geman (1984). In joint work with Peter Green and Bruce Porteous, we have had some success in reproducing lines (hedges and roads) in synthetic aperture radar (*SAR*) images.

5. SOME COMPLICATIONS

5.1. *Parameter Estimation*

We now consider the estimation of any unknown parameters θ and ϕ in f and p, respectively; θ and ϕ may be vectors. There are two distinct problems. In the first, it is required to estimate θ and ϕ from training data alone and to use the estimated values in subsequent reconstruction. In the second, no training data are available and it is necessary to estimate θ and ϕ as part of the restoration procedure. These two extremes may occur in combination: for example, training data may be available for θ but may not be relevant for ϕ.

5.1.1. *Estimation from training data*

Suppose the training data comprise records y for a relevant known scene x. Then, in principle, the method of maximum likelihood can be applied to obtain estimates $\hat{\theta}$ and $\hat{\phi}$, these being the values of θ and ϕ which maximize $l(y \mid x;\ \theta)$ and $p(x;\ \phi)$, respectively. The factorization (1) ensures that, apart from possible complications of multivariate y_is, $\hat{\theta}$ can be calculated in a straightforward manner. However, as regards ϕ, maximum likelihood estimation is computationally intractable, except for the Markov mesh models of Section 2.4 and simple first-order two-colour Ising models. In general, it is the constant of proportionality in $p(x;\ \phi)$ which cannot be evaluated.

A simple alternative to maximum likelihood estimation for local Markov random fields is provided by the "coding method" (Besag, 1974). This is a procedure in which the estimate $\hat{\phi}$ is chosen to maximize the conditional likelihood,

$$\prod_{i \in M} p_i(x_i \mid x_{\partial i};\ \phi),$$

where M is a (maximal) set of pixels, such that no two are neighbours. Different coding estimates can be combined to produce an average $\hat{\phi}$. For further details, see Besag (1974), and for examples on quite complex M.r.f.s, Cross and Jain (1983).

However, a neater and more efficient procedure is to adopt maximum "pseudo-likelihood" estimation (Besag, 1975), in which the estimate $\hat{\phi}$ is the value which maximizes the corresponding product over *all* pixels $i \in S$; in practice, one might exclude boundary pixels from the product because of the added artificiality of the model there. Of course, the product is not a genuine likelihood, except in the trivial case of independence. For some further discussion and an example, see Besag (1977b, 1978).

5.1.2. *Estimation during ICM*

When reconstruction and parameter estimation must be carried out simultaneously, we propose the following iterative procedure:

1. Obtain an initial estimate $\hat{x}$ of the true scene x^*, with guesses for θ and ϕ, if necessary.
2. Estimate θ by the value $\hat{\theta}$ which maximizes $l(y \mid \hat{x};\ \theta)$.
3. Estimate ϕ by maximum pseudo-likelihood on the current $\hat{x}$ to obtain a new $\hat{\phi}$; that is, choose $\hat{\phi}$ to maximize

$$\prod_{i \in S} p_i(x_i \mid x_{\partial i};\ \phi),$$

with the possible exclusion of boundary pixels from the product.

4. Carry out a *single* cycle of *ICM* as in Section 2.5, based on the current $\hat{x}$, $\hat{\theta}$ and $\hat{\phi}$, to obtain a new $\hat{x}$.
5. Return to 2 for a fixed number of cycles or until approximate convergence of $\hat{x}$.

Little is known of the convergence properties of the above procedure, particularly in multi-parameter estimation, but limited experience thus far seems encouraging. Note that if, as recommended for *ICM*, spatial information is ignored in the initial reconstruction, then $\hat{\phi}$ in stage 1 is not required; in the example below, this is also true of $\hat{\theta}$. Of course, if an external estimate of θ is available, perhaps from training data, then this can be used throughout reconstruction and stage 2 omitted.

5.1.3. *Example* 3

Fig. 3a represents the first 64 rows and last 64 columns of Fig. 2a, in which adjacencies are less contrived than in the scene as a whole. Records were produced as in Section 3 by superposition of univariate Gaussian noise with variance $\kappa = 0.36$. The initial reconstruction, based on the closest mean to each record, produced 32% misclassification; see Fig. 3b. With the correct value of κ and with $\beta = 1.5$ throughout, *ICM* gave a 2.1% misclassification after six cycles; with gradually increasing β over the first six cycles, this improved to 1.2% on the eighth cycle, as in Fig. 3c. Note that the majority-vote version of local reconstruction, with sufficiently large β, resulted in 9.3% misclassification after six iterations and 8.6% on convergence after 14 iterations: all genuine small groups of pixels were eliminated.

For κ known but β estimated as in Section 5.1.2, the eventual error rate was 1.1% after eight iterations (Fig. 3d), with $\hat{\beta} = 1.80$; when estimating both β and κ during reconstruction, the error rate again settled at 1.2%, now with $\hat{\beta} = 1.83$ and $\hat{\kappa} = 0.366$. All the above computations used a raster scan with asynchronous updating, although the programs were written in APL. Results were almost identical using synchronous updating at the end of each cycle.

The above results have been reproduced qualitatively on other examples but as yet we have no experience of more complicated situations. An interesting consequence of estimation in $p(x; \phi)$ is that, during the early stages of restoration, there is a gradual increase in the strength of the field as already recommended. This effect is enhanced by the estimation of κ. However, it is well known that mixture distributions must be handled with care: for some warnings, in the context of image processing, see Titterington (1984b).

5.2. *Pixel Overlap*

Suppose that each record Y_i, for a given scene x, depends not only on x_i itself but also on the x_j's of a known set vi in the vicinity of pixel i. Then the conditional density of Y_i given x can be written $f_i(y_i \mid x_i, x_{vi})$: we still assume the Y_i's are conditionally independent given x. For a rectangular array, v_i might comprise the eight nearest pixels to pixel i, with suitable modification at the boundary of S. However, note that, as distinct from the choice of neighbourhood ∂i for each pixel, there is no essential requirement for i to belong to vj if j belongs to vi.

In order to update the colour of pixel i on any cycle of *ICM*, it is again necessary to choose $\hat{x}_i$ to maximize $P(\hat{x}_i \mid y, \hat{x}_{S\setminus i})$, where $\hat{x}_{S\setminus i}$ denotes the current colouring elsewhere; that is, to choose $\hat{x}_i$ to maximize

$$f_i(y_i \mid \hat{x}_i, \hat{x}_{vi}) p_i(\hat{x}_i \mid \hat{x}_{\partial i}) \prod_{j \in \rho i} f_j(y_j \mid \hat{x}_j, \hat{x}_{vj}),$$

where $\rho i = \{j : i \in vj\}$. The additional computational burden is light but of course reconstruction will be less successful than in the absence of pixel overlap.

5.3. *Conditionally Dependent Records*

The assumption in Section 2.1 that records are conditionally independent, given the true scene, can be relaxed, particularly in the case of Gaussian records. Thus, Kittler and Föglein (1984a, b), Kittler and Pairman (1985) and Kiiveri and Campbell (1986) adopt a local auto-Normal formulation for the conditional distribution of Y given x and extend this to the case of

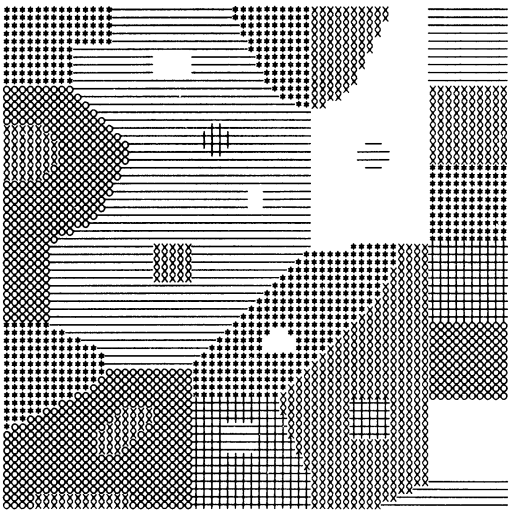

Fig. 3a. True six-colour scene: 64 × 64.

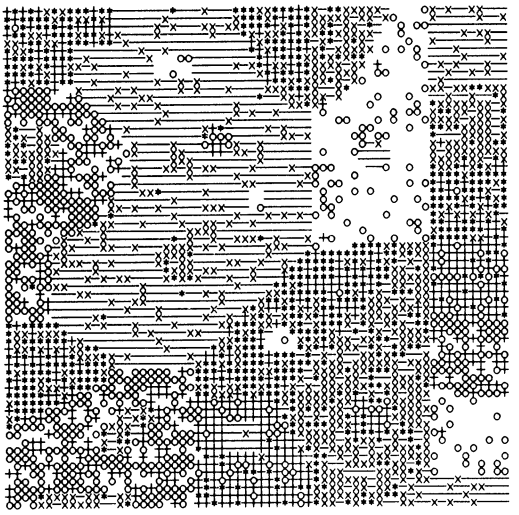

Fig. 3b. Maximum likelihood classifier: 32% error rate.

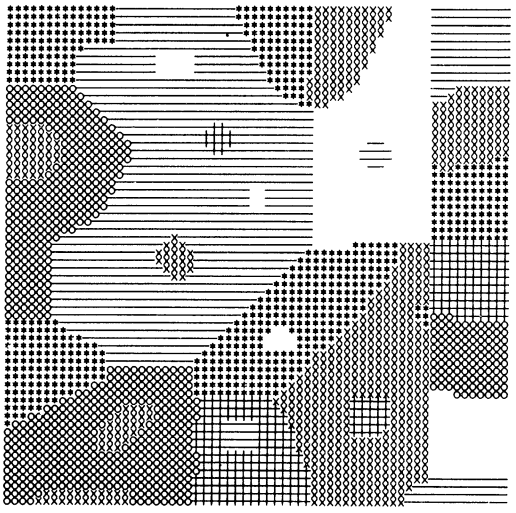

Fig. 3c. ICM reconstruction with $\beta \uparrow 1.5$: 1.2% error rate.

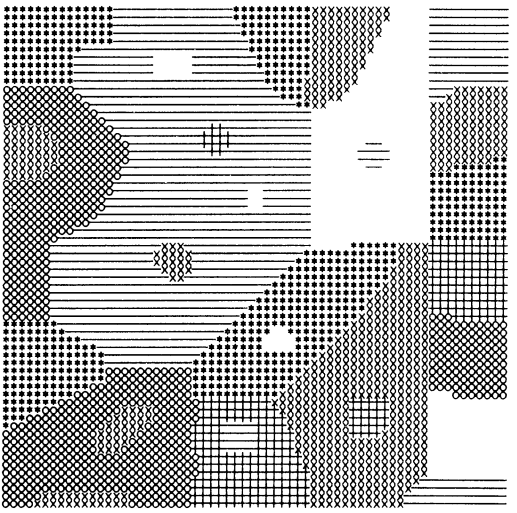

Fig. 3d. ICM reconstruction with β estimated: $\hat{\beta} = 1.80$: 1.1% error rate.

multi-component records. The results are applied to the classification of land usage and of clouds, from satellite data.

5.4. *Block Reconstruction*

In Section 2.6, it was proposed that, instead of choosing $\hat{x}_i$ to maximize $P(\hat{x}_i \mid y, \hat{x}_{S\setminus i})$ at any stage of reconstruction, one might choose $\hat{x}_B$ to maximize $P(\hat{x}_B \mid y, \hat{x}_{S\setminus B})$, where B represents a small block of pixels in the vicinity of pixel i. In particular, suppose that the Bs form 2×2 blocks of four, these being addressed in a raster scan with overlap allowed between successive blocks. At each stage, the block in question must be assigned one of c^4 colourings, based on four records and, in the case of second-order neighbourhood, 26 direct and diagonal adjacencies. With Assumptions 1 and 2 of Section 2.1 and $\{p(x)\}$ as in (6), this reduces to

choosing $\hat{x}_B$ to maximize

$$\left\{\prod_{i \in B} f(y_i \mid x_i)\right\} \exp(\beta \times \textit{number of like adjacencies}). \tag{17}$$

The computational burden is much heavier than before and the illustrative example below is restricted to $c = 2$ colours, though special purpose algorithms could of course be written. Note that parameter estimation can be incorporated as in Section 5.1.

Fig. 4a shows an 88 × 100 hand-constructed scene, designed specifically to contain some awkward features. Univariate Gaussian records were generated as before but with $\kappa = 0.9105$, giving an initial 30% misclassification using the linearly nearest mean; see Fig. 4b. Several runs of local reconstruction were carried out using (7), with pixels individually updated in a raster scan. The most basic version, with $\beta = 1.5$ in each of six cycles of restoration, gave a rather disappointing error rate of 9% but this improved to 6.2% using eight iterations and the previously recommended practice of gradually increasing β from 0.5 to 1.5 over the first six. In the third run, β was estimated after each cycle, as in Section 5.1, and this resulted in a further small gain to 5.7%; see Fig. 6c. Note here that although changes in reconstructions after the fourth cycle (error rate 5.6%) were very slight (1.5% of pixels), the maximum pseudo-likelihood estimates were altered radically. Specifically, $\hat{\beta} = 1.3$, 1.9 and ∞ after 4, 6 and 8 cycles, respectively! Of course, such estimation takes direct account only of the *local* properties of the current scene and it happens here that this is eventually of the majority-vote form. Incidentally, if pixels are synchronously updated at the end of each cycle, rather than in a raster scan, the above three versions of local reconstruction produce error rates of 6.1, 5.7 and 5.3%, respectively; furthermore, in the last case, the eventual estimate of β is 1.4, compared with the value 0.9 obtained from the true scene. This reflects the more forgiving nature of synchronous updating. However, Fig. 6c has been retained for comparability with block reconstruction, which also uses a raster scan.

Fig. 4d shows the result of applying six cycles of (17) with $\beta = 1.5$: the distortion is severe and there is a 13% error rate. Several variants were tried, including an assortment of βs, both constant and gradually increasing, and downweighting for diagonal adjacencies. Fig. 4e shows the best result, with $\beta = 0.5$ and constant: convergence occurred at the third iteration with an error rate 5.6%. The findings support the idea that quite different values of β may be optimal for different block sizes, as patterns are mimicked at different scales. The same idea extends to parameter estimation in $p(x; \phi)$: thus, maximum pseudo-likelihood estimation, which concentrates on the boundaries between (unordered) colours, may be inappropriate for m.a.p. reconstruction.

6. CONCLUDING COMMENTS

In this paper, we have pursued *ICM* as a means of restoring degraded scenes. We have considered several artificial examples in which degradation arose through the simple superposition of independent Gaussian noise on the colour labels, though the technique applies to any situation in which the records are conditionally independent, given the true scene, and to some in which they are not, provided the corresponding distribution is of known parametric form. Thus, in nuclear medicine, it will often be appropriate to assume that the records have independent Poisson distributions, the means of which, or some functions thereof, we wish to estimate; whereas, in applications involving synthetic aperture radar, theoretical considerations suggest that degradation occurs multiplicatively and in accordance with a Rayleigh distribution (Bush and Ulaby, 1975).

In conclusion, it may be useful to briefly review the philosophy of our approach and to pose some questions. We began, in Section 2, by adopting a strict probabilistic formulation with regard to the true scene and to the records it generates. This led to global criteria for reconstruction, as described in Sections 2.3 and 2.4. However, we then abandoned these in

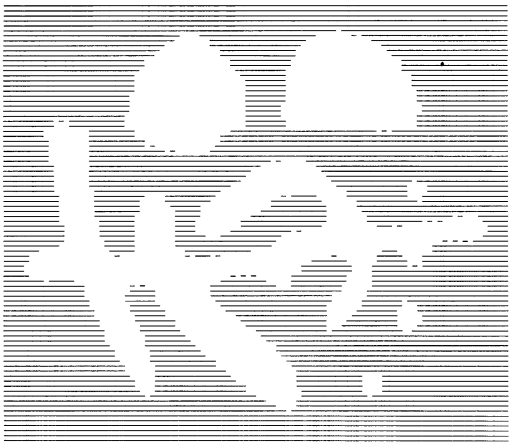

Fig. 4a. True two-colour scene: 88 × 100.

Fig. 4b. Maximum likelihood reconstruction: 30% error rate.

Fig. 4c. ICM reconstruction with β estimated: $\hat{\beta} = \infty$: 5.7% error rate.

Fig. 4d. ICM reconstruction on 2 × 2 blocks with $\beta = 1.5$: 13% error rate.

Fig. 4e. ICM reconstruction on 2 × 2 blocks with $\beta = 0.5$: 5.6% error rate.

favour of *ICM*, partly on grounds of computation and partly to avoid any unwelcome large-scale effects, which might otherwise be encouraged by $\{p(x)\}$. Such reconstruction may be plausible but no longer has a proper mathematical basis. Thus, one might legitimately question the restriction to genuine M.r.f.s in choosing the conditional distribution $p_i(x_i \mid x_{\partial i})$ at each pixel; after all, *ICM* itself does not require this. To take a very simple example, might one not choose the conditional odds of colour k at pixel i, given the colours at all other pixels, to be proportional to the number $u_i(k)$ of like-coloured neighbours, rather than $\exp\{\beta u_i(k)\}$ as in (6), even though this is strictly invalid? It is only a partial answer to suggest that adherence to genuine M.r.f.s removes some arbitrariness from the choice and may aid interpretation.

There are many other problems we have not addressed or which remain unresolved. Perhaps the most important concerns the practical relevance of Assumption 1 and its generalization to include pixel overlap. Another is the assumption that each pixel has a unique true colour, which is of course false in practice: thus, Kent and Mardia (1986) allow mixed pixels in their formulation, while Hjort and Mohn (1984) include a "doubt" class in their reconstructions. To continue: are one's misgivings about global reconstruction criteria justified? Is there any means of making valid inferential statements about reconstructions, given the crude nature of the spatial modelling? How should special features in the true scene be modelled? How should algorithms be compared, given that error rates have severe limitations and do not focus on colour boundaries?

In short, there is a vast number of problems, not just in the relatively narrow context of the present paper but in image processing and in pattern recognition as a whole, to which statisticians might usefully contribute.

ACKNOWLEDGEMENTS

I am grateful to the many people both in Durham and elsewhere with whom I have had helpful discussions. I particularly thank Peter Green and the organizers and participants in the 1985 Edinburgh Workshop on Statistics and Pattern Recognition, sponsored by the Society and supported by the Science and Engineering Research Council. The constructive comments of the referees were much appreciated.

REFERENCES

Abend, K., Harley, T. J. and Kanal, L. N. (1965) Classification of binary random patterns. *I.E.E.E. Trans. Inform. Th.*, **11**, 538–544.

Bartlett, M. S. and Besag, J. E. (1969) Correlation properties of some nearest-neighbour models. *Bull. Int. Statist. Inst.*, **43**, 191–193.

Besag, J. E. (1974) Spatial interaction and the statistical analysis of lattice systems (with Discussion). *J. R. Statist. Soc.* B, **36**, 192–236.

——(1975) Statistical analysis of non-lattice data. *The Statistician*, **24**, 179–195.

——(1976) Parameter estimation for Markov fields. *Princeton Univ. Dept. of Statistics Tech. Report* 108, Series 2, 38pp.

——(1977a) On spatial temporal models and Markov fields. *Proc. 10th European Meeting of Statisticians*, Prague, 1976, 47–55.

——(1977b) Efficiency of pseudolikelihood estimation for simple Gaussian fields. *Biometrika*, **64**, 616–618.

——(1978) Some methods of statistical analysis for spatial data. *Bull. Int. Statist. Inst.*, **47**, 77–92.

——(1983) Discussion of paper by P. Switzer. *Bull. Int. Statist. Inst.*, **50** (Bk. 3), 422–425.

Burch, S. F., Gull, S. F. and Skilling, J. (1983) Image restoration by a powerful maximum entropy method. *Comput. Vision Graph. and Image Process.*, **23**, 113–128.

Bush, T. F. and Ulaby, F. T. (1975) Fading characteristics of panchromatic radar backscatter from selected agricultural targets. *I.E.E.E. Trans. Geosc. Elect.*, **13**, 149–157.

Cross, G. C. and Jain, A. K. (1983) Markov random field texture models. *I.E.E.E. Trans. Pattern Anal. Machine Intell.*, **5**, 25–39.

Derin, H., Elliot, H., Christi, R. and Geman, D. (1984) Bayes smoothing algorithms for segmentation of binary images modelled by Markov random fields. *I.E.E.E. Trans. Pattern Anal. Machine Intell.*, **6**, 707–720.

Devijver, P. A. (1985) Probabilistic labelling in a hidden second order Markov mesh. *In 'Pattern Recognition in Practice II'*, Ed. E. Gelesma and L. Kanal. North-Holland.

Devijver, P. A. and Kittler, J. (1982) *Pattern Recognition: A Statistical Approach.* Englewood-Cliffs: Prentice Hall.

Fu, K. S. and Yu, T. S. (1980) *Statistical Pattern Classification using Contextual Information.* Chichester: Research Studies Press.

Galbraith, R. F. and Walley, D. (1976) On a two-dimensional binary process. *J. Appl. Prob.*, **13**, 548–557.

Geman, S. and Geman, D. (1984) Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *I.E.E.E. Trans. Pattern Anal. Machine Intell.*, **6**, 721–741.

Geman, S. and McClure, D. E. (1985) Bayesian image analysis: an application to single photon emission tomography. *Proc. Amer. Statist. Assoc.* Statistical Computing Section, 12–18.

Green, P. J., Jennison, C. and Seheult, A. H. (1985) Analysis of field experiments by least squares smoothing. *J. R. Statist. Soc.* B, **47**, 299–315.

Grenander, U. (1983) Tutorial in Pattern Theory. *Division of Applied Mathematics, Brown University.*

Gull, S. F. and Daniel, G. J. (1978) Image reconstruction from incomplete and noisy data. *Nature*, **272**, 686–690.

Hand, D. J. (1981) *Discrimination and Classification.* London: Wiley.

Hansen, F. R. and Elliott, H. (1982) Image segmentation using simple Markov field models. *Comput. Vision Graph. and Image Process.*, **12**, 357–370.

Haslett, J. (1985) Maximum likelihood discriminant analysis on the plane using a Markovian model of spatial context. *Pattern Recognition*, **18**, 287–296.

Hastings, W. K. (1970) Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**, 97–109.

Hjort, N. L. and Mohn, E. (1984) A comparison of some contextual methods in remote sensing classification. *Proc. 18th Int. Symp. on Remote Sensing of Environment*, Paris, 1–11.

Hummel, R. A. and Zucker, S. W. (1983) On the foundations of relaxation labeling processes. *I.E.E.E. Trans. Pattern Anal. Machine Intell.*, **5**, 267–287.

Kanal, L. N. (1980) Markov mesh models. In *Image Modelling.* New York: Academic Press.

Kay, J. W. and Titterington, D. M. (1986) Image labelling and the statistical analysis of incomplete data. *Proc. 2nd Int. Conf. Image Processing and Applications.* Conf. Publ. No. 265. London: Inst. Elec. Engrs., pp. 44–48.

Kent, J. T. and Mardia, K. V. (1986) Spatial classification using fuzzy membership models.

Kiiveri, H. T. and Campbell, N. A. (1986) Allocation of remotely sensed data using Markov models for spectral variables and pixel labels.

Kinderman, R. and Snell, J. L. (1980) *Markov Random Fields and their Applications.* Providence: American Mathematical Society.

Kirkpatrick, S., Gelatt, C. D. and Vecchi, M. P. (1983) Optimization by simulated annealing. *Science*, **220**, 671–680.

Kittler, J.and Föglein, J. (1984a) Contextual classification of multispectral pixel data. *Image and Vision Computing*, **2**, 13–29.

——(1984b) Contextual decision rules for objects in lattice configurations. *Proc. 7th Int. Conf. Pattern Recognition*, Montreal, 1984, 270–272.

Kittler, J. and Pairman, D. (1985) Contextual pattern recognition applied to cloud detection and identification. *I.E.E.E. Trans. Geosci. Remote Sensing*, 23, 855–863.

Pickard, D. K. (1977) A curious binary lattice process. *J. Appl. Prob.*, **14**, 717–731.

Pickard, D. K. (1980) Unilateral Markov fields. *Adv. Appl. Prob.*, **12**, 655–671.

Preston, C. J. (1974) *Gibbs States on Countable Sets.* Cambridge: University Press.

Rosenfeld, A. and Kak, A. C. (1982) *Digital Picture Processing* (2 vols). New York: Academic Press.

Rosenfeld, A., Hummel, R. A. and Zucker, S. W. (1976) Scene labelling by relaxation operations. *I.E.E.E. Trans. Syst. Man. Cybern.*, **6**, 420–433.

Saebo, H. V., Braten, K., Hjort, N. L., Llewellyn, B. and Mohm, E. (1985) Contextual classification of remotely sensed data: statistical methods and development of a system. *Report No.* 768, *Norwegian Computing Centre*, Oslo.

Skilling, J. (1984) The maximum entropy method. *Nature*, **309**, 748–749.

——(1986) Maximum entropy data analysis.

Strauss, D. J. (1977) Clustering on coloured lattices. *J. Appl. Prob.*, **14**, 135–143.

Switzer, P. (1980) Extensions of linear discriminant analysis for statistical classification of remotely sensed satellite imagery. *Math. Geol.*, **12**, 367–376.

——(1983) Some spatial statistics for the interpretation of satellite data (with Discussion). *Bull. Int. Statist. Inst.*, **50** (Bk. 2), 962–972.

Titterington, D. M. (1984a) The maximum entropy method for data analysis (with reply). *Nature*, **312**, 381–382.

——(1984b) Comments on paper by S. C. Sclove. *I.E.E.E. Trans. Pattern Anal. Machine Intell.*, **6**, 656–657.

——(1985) Regularization procedures in signal processing and statistics. *Proc. IMA Conf. Math. and Signal Processing*, Bath, 1985. Oxford: University Press.

Yu, T. S. and Fu, K. S. (1983) Recursive contextual classification using a spatial stochastic model. *Pattern Recognition*, **16**, 89–108.

DISCUSSION OF PROFESSOR BESAG'S PAPER

Professor D. M. Titterington (University of Glasgow): In recent years there has been some concern that mainstream statisticians have been usurped in some areas of research that are essentially statistical. As a result, groups such as econometricians, computer scientists and information technologists may be thought to have developed unjustifiably high profiles compared with the humdrum statistician. One such area is that of pattern recognition and image processing, generally thought to be the prerogative of "engineers", but which abounds in the raw material for research by statisticians. The potential for involvement is reflected in the Edinburgh workshop mentioned by Professor Besag and, more specifically, by his admirable paper.

It is important to answer two main questions:

(A) What can statisticians offer to the field?
(B) Do the results of their efforts lead to *useful* methods?

The most important aspect of the answer to (A) is "experience in statistical modelling and analysis". A crucial feature of tonight's paper is the proposal of a joint probability function

$$p(x, y; \theta, \phi) = l(y \mid x; \theta)p(x; \phi) \qquad (*)$$

and the adoption of particular forms for l and p on the right-hand side. Given (θ, ϕ), maximisation with respect to x provides $\hat{x}$, the m.a.p. estimate of x. ($p(x, y)$ and $p(x \mid y)$ are equivalent in this respect.) It is also suggested, roughly speaking, that "maximum likelihood" estimates of (θ, ϕ) can be obtained from

$$p\{\hat{x}(\theta, \phi), y; \theta, \phi\}.$$

In fact this would be more reliably achieved by maximizing the "marginal"

$$p(y; \theta, \phi) = \sum_x p(x, y; \theta, \phi),$$

with the x regarded as missing data. However, even use of the *EM* algorithm in this context becomes very complicated, because of the effect, on the *E*-step, of the marginal (prior) correlations among the components of x. I must admit to a preference for the nomenclature "conditional density" and "marginal density" for the factors of (*), rather than the Bayesian-flavoured "likelihood" and "prior". As a result, I can regard the y as mixture data with the complication that the underlying (unobservable) classification variables making up x are not independent; see Baum *et al.* (1970) for a version of *EM* for the one-dimensional case, in which the components of x follow a Markov chain, and Geman and McClure (1985) for a variation of *EM* which is based on a Monte Carlo treatment of the *E*-step.

Two important difficulties arise in connection with the model. Firstly, in spite of the effort expended in describing the Markov random field models for $p(x)$, it is admitted in the paper that they are probably globally unacceptable. Secondly, calculation of $\hat{x}$, as defined, is rejected as being impracticable. To put it bluntly, it could be said that the stylish modelling turns out to be a means to the end of proposing, in the form of the *ICM* method, a practically feasible, if rather *ad hoc* algorithm. Against this, of course, the model may be more acceptable locally than it is globally and I wonder whether there is useful cross-fertilisation with O'Hagan's (1978) work on locally linear regression.

The need for *ad hoc* modification leads me to view the procedure underlying the definition of $\hat{x}$ on a basis more informal than the Bayesian one. Note that, with mention of (θ, ϕ) suppressed,

$$\log p(x, y) = \Phi_1(y, x) + \Phi_2(x),$$

where Φ_1 represents fidelity of x to the data and Φ_2 is a function measuring local smoothness of the image: c.f. formulae (7) and (8). The M.r.f. assumptions happen to provide sensible forms for Φ_2 but, beyond that, one does not need to believe in the M.r.f. as a model. This approach to the definition of $\hat{x}$ is formally equivalent to ridge regression (note (14) and (15)) and other "nonparametric" regularisation procedures including the maximum entropy method (Titterington, 1985). The parameter β within Φ_2 is the smoothing parameter.

If this new interpretation of the problem is accepted, the following questions arise.

(i) Is it possible, or worthwhile, to choose β by some technique, such as crossvalidation, that is distinctly non-Bayesian?

(ii) In the spirit of variable kernel density estimation, is it appropriate, or worthwhile, to choose different values of β in different parts of the image?

The word "worthwhile" leads me on to question (B).

In whatever framework the methods are imbedded, for real images the weight of calculations and complexity of methodology will have to be traded off against the quality of the result and the practical requirements. This prompts the question of what constitutes a good image. Pure error rate can be very misleading, as can any single criterion of fit between the restored and true images. Some criteria might be very specific, such as the ability to restore small features such as the peninsula in Fig. 1. (That a *smoothing* procedure such as *ICM* manages to recover such details is impressive and perhaps surprising.) A crucial factor may be whether the restored image requires to be interpreted automatically or simply to be examined by a human being. Many papers report different restorations of the same "standard" photographs with, frankly, little to choose visually among them.

It will however be invaluable to carry out meaningful, comparative studies of the performances of a variety of restoration techniques, using different performance criteria on different types of image. It will be important to bear in mind the computational complexity involved. For one thing, gross numerical investment is not likely to pay off. For another, we presumably want our methods to be useful for large-scale images and, ultimately, for real-time film processing: this will require virtually instantaneous restoration and will favour reliable, simple techniques.

In conclusion, and in spite of my apparent disappointment that the formal modelling approach has to be diluted to become workable, I believe that the ideas of the paper are more secure than most of the other relaxation labelling techniques (Hummel and Zucker, 1983) and Professor Besag is to be applauded for leading UK statisticians into this important area. Maybe his technique's title can be extended eponymously to that of Iterated Conditional Besag Modes to reflect, by its initials, his international and explosive impact on the field!

I am very pleased to be able to propose the vote of thanks.

Dr J. Kittler (University of Surrey): We have been investigating the role of iterative conditional modes (using Professor Besag' terminology) in image segmentation and classification for some time now. Prof Besag gives an excellent overview of this approach, the motivation behind it, its relationship with other methods and last but not least its application potential. We consider one of the crucial issues if not the crucial one to be how to determine an appropriate model for the task in hand. Prof Besag addresses two aspects of this issue. One deals with the question of selecting a model type that would faithfully represent the properties and relationships of pixel labels (colours) in the uncorrupted image. In terms of the Gibbs distribution this corresponds to specifying suitable functional forms of the G-functions.

The second and closely related aspect is concerned with the problem of estimating the parameters of the G-functions and we have been given some suggestions as to how the parameter estimation problem might be approached.

The third aspect, only touched upon in Professor Besag's paper, relates to model structure. While the general form of the Gibbs distribution is a vehicle for describing any order of interactions between pixels, a successful modelling of the true scene may be entirely dependent on the model structure first being correctly identified. Such a statement would be met with little controversy in other modelling contexts and it would be surprising if it were not pertinent in image modelling.

Although the Gibbs distribution permits an extensive freedom in specifying valid markov random fields, our experience indicates that the order of neighbourhood, the order of interactions and the structural modes of the Gibbs distribution are intimately related and their relationship imposes some constraints on image models. Thus one cannot for instance adopt the second order neighbourhood in conjunction with pairwise interactions without implicitly assuming that the corresponding model only approximates the true interactions between pixel labels in the image. Without a compromise, pairwise interactions can be considered for the first order neighbourhood only, while interactions between triplets and pairs are appropriate for the non-lattice type neighbourhood of Fig. D1, quartets for the second order neighbourhood, etc. Now the existence of a close relationship between the highest order of interactions and the type and order of the neighbourhood is not surprising as it follows directly from the Hammersley and Clifford theorem. What is more unexpected, at least to us, is that each type of contextual neighbourhood appears to admit a specific family of Gibbs distributions characterised by a distinct structure. For instance the second order neighbourhood does not give rise to triplets and the pairwise interactions there are relevant in orthogonal directions only. In the case of the edge model of Geman and Geman with the neighbourhood as depicted in Fig. D2, the Gibbs distribution contains two fourth order modes but no third and second order interactions.

With these observations in mind we may then ask a number of pertinent questions. What effect does the use of admissible rather than general structure have on modelling. How well and to what extent

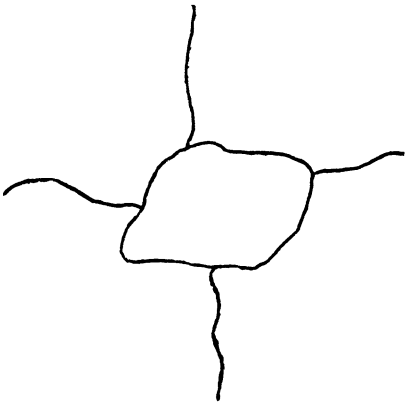

Fig. D1

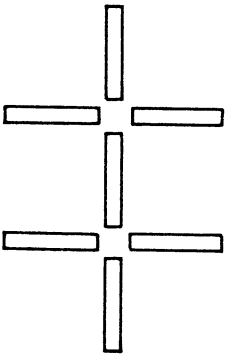

Fig. D2

can high order interactions be approximated by low order ones. If low order approximations are found adequate, why is is that one can dispense with the higher order terms. Furthermore, is the use of higher order neighbourhoods in conjunction with low order interactions superfluous.

The existence of these and other open questions identified elsewhere does not in any sense diminish the contributions made in Professor Besag's paper to the subject of Markov random fields and their application to image restoration and analysis. This lucid paper is not only illuminating but will no doubt stimulate in the true Besag tradition a lot of theoretical and experimental work in the future. I am very pleased to second the vote of thanks.

The vote of thanks was passed by acclamation.

Drs D. M. Greig, B. T. Porteous and A. H. Seheult. (University of Durham): We would like to congratulate our colleague, Julian Besag, on a stimulating paper that, hopefully, will encourage other statisticians to work on the challenging practical and theoretical problems in image processing.

We show that a m.a.p. estimate of any two-colour scene can be found exactly using the labelling algorithm of Ford and Fulkerson (1956) for finding the maximum flow in a certain network. Possible extension to more colours is being investigated.

In the binary case, for any noise distribution, equation (8) can be written

$$-\sum_{i=1}^{n}[x_i \log f(y_i \mid x_i = 1) + (1 - x_i) \log f(y_i \mid x_i = 0)]$$

$$-\beta \sum_{1 \leqslant i < j \leqslant n}\sum a_{ij}[x_i x_j + (1 - x_i)(1 - x_j)]$$

where $a_{ij} = 1$ if i and j are neighbours and zero otherwise. This simplifies to a quadratic function of the form

$$p + \sum_{i=1}^{n} p_i x_i - \sum_{i=1}^{n}\sum_{j=1}^{n} p_{ij} x_i x_j \tag{1}$$

where $p_{ij} \geqslant 0$.

Picard and Ratliff (1975) show that minimising (1) subject to $x_i = 0$ or $x_i = 1 (i = 1, \ldots, n)$ is equivalent to finding the minimum cut, or equivalently the maximum flow, in a network. Arc capacities from the source, to the sink and between internal nodes are determined by the p_i and p_{ij}. For a binary scene, internal nodes correspond to pixels and are connected if the pixels are neighbours; capacities of all internal connections are β. If the log-likelihood ratio $\lambda_i = \log\{f(y_i \mid x_i = 1)/f(y_i \mid x_i = 0)\}$ is positive then there is a connection from the source to pixel i with capacity λ_i; otherwise, the connection is to the sink with capacity $-\lambda_i$. Thus pixel i is connected to the source if and only if its max. likelihood classification is $x_i = 1$.

In a m.a.p. estimate, $x_i = 1$ if pixel i and the source are on the same side of the minimum cut; otherwise, $x_i = 0$. Our implementation of the Ford-Fulkerson algorithm exploits the structure of the present problem, but generalises to include any neighbourhood system defined on cliques of size two.

Fig. D3a. M.a.p. estimate: $\beta = 1/3$, 5.0% error rate.

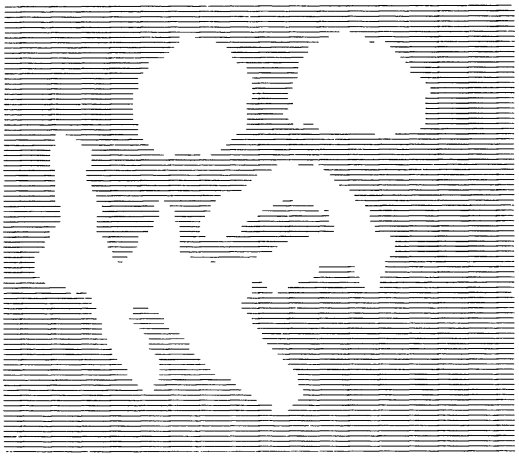

Fig. D3b. M.a.p. estimate: $\beta = 1/2$, 6.4% error rate.

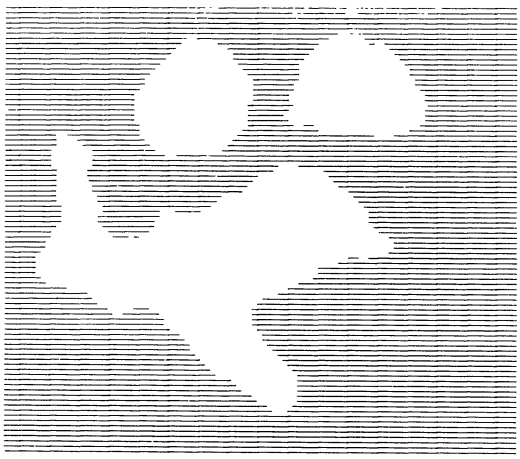

Fig. D3c. M.a.p. estimate: $\beta = 2/3$, 10.2% error rate.

Fig. D3d. Simulated annealing: $\beta = 1/3$, 5.5% error rate.

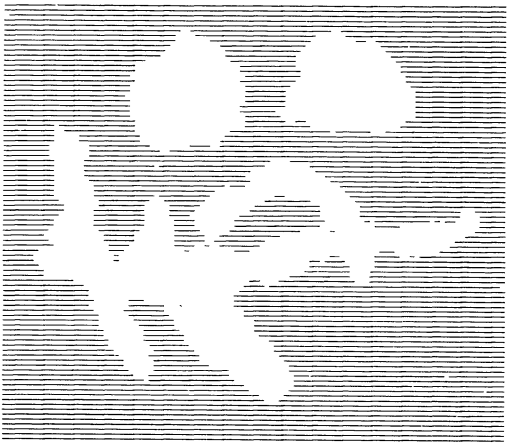

Fig. D3e. Simulated annealing: $\beta = 1/2$, 5.8% error rate.

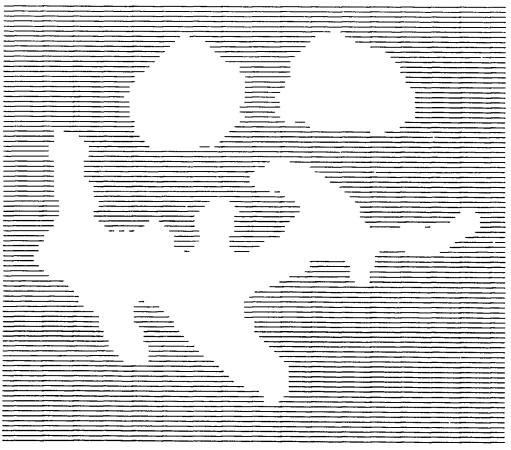

Fig. D3f. Simulated annealing: $\beta = 2/3$, 7.6% error rate.

For different β's, Fig. D3 displays m.a.p. estimates of the binary image considered in Section 5.4. We used a different realisation of the same noise distribution so results are not directly comparable with Professor Besag's Fig. 4.

Fig. D3 also displays the 750th iterations of simulated annealing with temperature schedule $T_k = 2/\log(1+k)$. These were obtained using part of an image processing package developed by Peter Green.

Inspection of Fig. D3 suggests that;

1. Relative to ICM, a smaller value of β may be more appropriate for m.a.p. estimation.
2. For a given noise level, simulated annealing "converges" to m.a.p. more rapidly for smaller values of β.

The CPU times used per image were roughly 7 minutes for simulated annealing and one hour for m.a.p. Following a suggestion of Frank Kelly and Peter Green the m.a.p. times have now been reduced to around 9 minutes.

Dr Peter Clifford (Oxford University): I would like to add my congratulations to those of the other contributors on a paper in the best traditions of this Society. By this I mean that the paper is centred about an important body of data. On the whole statisticians have not involved themselves wholeheartedly in the new problems of data analysis which arise when data sets are large and the parameter space has correspondingly high dimension. Such data are typical in the physical and engineering sciences. The analysis of such data has been left to engineers, computer scientists and to an increasing extent the new breed of applied mathematicians. If the statistical community is to retain its importance as a clearing house for new ideas in the subject, then statisticians must be prepared to go out, track down the data and get involved. What I am saying is that the analysis of dirty pictures is not the only way of getting your hands dirty.

To turn to a specific point, I am concerned about the assumption that the smoothing parameter β is constant over the image. Of course one may have very strong prior beliefs in favour of this assumption, but in general a more flexible approach is to model the smoothing parameter itself as a realisation of a spatial Markov field perhaps with pairwise interaction given by the Geman and McClure form of ϕ_{ij}. The conditional probability of colour k given not only the colours on other sites but also the value of the smoothing parameter process is then proportional to $\exp\{\alpha_k + \beta_i u_i(k)\}$ in the notation of (12). The pair $\{x^*_i, \beta_i\}$ is then a (bivariate) Markov field and can be reconstructed as a bivariate process by the methods described in Professor Besag's paper.

Dr P. J. Green (University of Durham): I would like to thank Professor Besag for introducing me to this subject, and congratulate him on an impressive paper.

All forms of image reconstruction based on equation (5) described in this paper are appealing both for reasons of computational economy, and because they seem to involve only local properties of the prior random field. Since certain obvious features in real images have *global* structure, however, it is interesting to examine whether global assumptions can be integrated into the author's elegant use of *locally*-dependent Markov random fields. I will refer only to the classification problem, with which the paper is mainly concerned, but the following points may be useful in a variety of other image-processing tasks.

Given a locally-dependent Markov random field $p(x)$, either the prior or the posterior field given in equation (3), let $D(x)$ be any non-negative random variable, measuring the extent to which the realisation x departs from some ideal property. Now form the modified field $p^*(x) = e^{-\sigma D(x)} p(x)/E(e^{-\sigma D})$ for some positive parameter σ. This shrinks the prior towards the desired property, and may be conceptually regarded as a rejection method: x from $p(x)$ is accepted with probability $e^{-\sigma D(x)}$. The resulting $p^*(x)$ is an *arbitrary* positive distribution on Ω, but this generality is of no interest: the point is that some such $p^*(x)$ are "locally realisable".

In the context of classification, one might take

$$D(x) = \sum_{k=1}^{c} |n_k(x) - \bar{n}_k|^{\lambda}$$

where $\{\bar{n}_k\}$ are specified target values. With the power $\lambda = 2$, for example, this modification has the simple effect on equation (12) of replacing α_k by $\alpha_k - 2\sigma(n_k(x_{S\backslash i}) - \bar{n}_k)$, involving the current frequencies. A heuristic argument suggests that for moderate sized images, say 100 square, even σ as small as 0.001 has a dramatic effect on reducing the expectation of D. This is borne out by simulations, which are qualitatively similar to Cross and Jain's simulations of the conditional prior, corresponding to $\sigma \to \infty$.

The implications of this are two-fold. Firstly, it is clear that if p^* is used in equation (5), starting from an initial x with colour frequencies close to target, then in a few iterations, the modification can make practically no difference to reconstruction using *ICM*. This is further evidence of the robustness of *ICM* to certain properties of the prior – it is equally well justified by p^*, the shrink prior yielding globally more desirable realisations, as by the original p. Secondly, in any method of restoration using (5) that converges more slowly, one could contemplate actually using this modified prior (the extra computing effort is trivial). This might be advantageous in examples where the spatial context acts differentially on the colours.

This agrees with experiments in which one colour appears only in rather fine features in the true scene. In *ICM* reconstructions, this colour can be seriously under-represented, increasing the pixel misclassification rate. This effect is even worse in the approximation to the maximum marginal probability classification obtained by accumulating modes over 30 or so sweeps. A much reduced error rate can be obtained by shrinking the prior towards the *MLC* colour frequencies.

Professor Philip J. Brown (University of Liverpool): In Section 5.1.1. I was struck by the phrase "*relevant* known scene x". Now calibration provides a somewhat more general model than the discrimination model used in tonight's paper for Y given x. Given training data Y (possibly multivariate) for fixed x, we are able to estimate the conditional distribution of Y given x, where x may be discrete or continuous, univariate or multivariate. Sometimes further *random* X are measured and we are then able to construct the conditional distribution the other way around, namely $X^* \mid y$ for a new observed picture y. Such a process relies on X being exchangeable with X^*, the true scene for the observed picture and inference for controlled calibration, in the absence of such random X may be much degraded, see Brown (1982). One may substitute a prior distribution of X^* and augment it if one has random X and call it a posterior distribution, or indeed, ignore many of its implications as with tonight's M.r.f. prior. It is, however, at least sensible that some empirical evaluation of the assumptions of the prior albeit local be addressed for *real* presented scenes. The prior is being used as an ingredient with strong local implications and surely, for example, it is an empirical question as to whether p or $\ln p$ is proportional to the number of like-coloured neighbours.

In Landsat satellite data and other applications Y are multivariate. Such multivariate observations can provide contradictory information about the underlying x, see Brown and Sundberg (1986, 1987). Such areas of the picture where the spatial prior is more influential need highlighting and is an aspect of the need for some attention to confidence as well as point estimation. Confidence considerations are also paramount when one wishes to estimate the area covered by a 'colour' as for example in the use of Landsat for forestry inventories, see for example Poso, Häme and Paananen (1984). Incidentally, it is evident in such Landsat work that the true scenes in training data are often not so easy or feasible to specify accurately as might be idealistically supposed.

Despite such criticisms, I have found tonight's paper extremely thought provoking and an impressive in-road into problems in the area.

Professor B. W. Silverman (University of Bath): I want to make a few remarks comparing the use of the m.a.p. estimator to the use of the simulation procedure suggested by Geman and Geman (1984) referred to at the end of Section 2.4 of tonight's paper. It is indeed plausible that the m.a.p. estimator will give the same result as the simulation procedure for the colour of a single pixel. However, any property of the picture that is not a linear function of the pixel values will not be estimated in the same way from the simulation model as from the m.a.p. estimator. The reason for this is that the operation of taking the posterior mode may have a drastic smoothing effect on the posterior distribution; if the picture X has a Gaussian distribution, so that m.a.p. and posterior mean coincide, if $\hat{x} = E_{\text{post}}(X)$, and if Ψ is any non-linear property then the simulation-based estimator $E_{\text{post}}\Psi(X)$ will differ from the m.a.p. estimator $\Psi(\hat{x})$. For measures Ψ of the fuzziness of X the inequality may be dramatic and indeed it is possible, for pictures defined on continuous space rather than just on pixels, to have $\Psi(X)$ infinite with probability one and $\Psi(\hat{x})$ finite. See Sections 6.1 and 7.3 of Silverman (1985).

This difficulty by no means undermines the importance of tonight's paper, but it is important to be aware of it in model-building. It means that local characteristics of the picture, in particular those relating to textures, built into the underlying model may in fact have little or even no chance of appearing in the m.a.p. and *ICM* estimators no matter what data are actually observed. This intuition is based on models where the value taken by the X_i are not confined to a discrete set, and therefore is more likely to be

applicable in a 256 grey-level picture than in a black/white one. Nevertheless, I believe it is an important area for detailed thought and work.

Dr C. D. Kershaw (Rothamsted Experimental Station): I would like to thank Professor Besag for presenting a very interesting paper. It is clear that improving discrimination accuracy by reference to local context is an important area for research.

I have recently used linear discrimination, and also iterative conditional modes (*ICM*) to classify pixels in a 2 km × 2 km scene of an area in the north of Cumbria. The data have been collected by a high-flying aircraft carrying a scanner similar to that on the French SPOT satellite. The pixel size was 20 m × 20 m. With this resolution it was possible to identify clearly field boundaries and roads. I identified 19 sites in the scene as being in one of seven land uses. I used spectral intensities from pixels at these sites as training data for discriminating all pixels within the scene. Intensities used were from green and near infra-red bands.

The allocation of pixels to land uses were tidier after 5 iterations of *ICM* (with $\beta = 0.75$) than for linear discrimination. Small clusters of pixels that were classified into a different use than most of their neighbours were usually reclassified. With linear discrimination, percentages correctly classified within the training set were over 95 per cent for all except grass and rough grazing. Grass and rough grazing had 62 and 58 per cent of pixels correctly classified. Most misclassifications were due to confusion between these two uses. The percentages correctly classified were virtually unaltered after the iterations of *ICM*. Grass and rough grazing had 64 and 58 per cent, respectively, of pixels correctly classified. Although *ICM* iterations have produced a more plausible allocation for pixels, the classification of pixels, on the whole, have not been dramatically altered.

Looking in more detail, it was apparent that the variabilities of spectral intensities were different for each land use, and that the variabilities between sites within the same land use were usually much greater than the within-site variability. Further calculations applying quadratic discrimination, classifying each site in the training set using remaining sites as a basis for discrimination, indicated that the percentage correct classification rates within the training set could be very misleading measures of accuracy. For example, one site was completely wrongly classified when classified using remaining sites.

These are the only data at which I have looked in such detail, but I suspect that many other satellite scenes of agricultural land have similar patterns of variability. In these circumstances, *ICM* will not significantly improve classification accuracy because smoothing out local anomalies by reference to neighbouring pixels will not work well where correlations across whole sites are high and there are large between-site variations. Automatic discriminators are becoming increasingly available in image analysis packages. Introduction of ICM in these packages could be dangerous as it will usually produce an allocation of pixels which is more convincing than that produced by a technique such as linear discrimination, but may not be greatly improving actual numbers of pixels correctly classified.

Dr Frank Critchley (University of Warwick): It is a pleasure to add my congratulations to the author on tonight's paper. The spirit of my comments is to underline several open questions. Some possible solutions were mentioned at the meeting. Limited space precludes repeating my suggestions here.

The general problem posed is the following. Estimate n colours x^* from:

(a) the observations y,
(b) *known* functions $f(.|.)$ and $p(.)$,
(c) *pre-determined* neighbourhoods
and (d) the *simplifying Assumptions* 1 and 2

Note that (b), (c) and (d) are all rather strong assumptions. The major difficulty is that n is very large, perhaps 2^{16}. In particular:

(1) *Uncertainty* about x^* exists. That is, there are many x's nearly as good as $\hat{x}$. The results of Critchley and Ford (1984, 1985) suggest that when we express our uncertainty about any unknown parameters in f or p, the inevitable increase in the uncertainty about x^* can be quite dramatic. Overall, how are we to represent this uncertainty? This problem is non-trivial, even if $\hat{x}$ is unique, (*cf.* interval estimation of an entire density or regression function, rather than of its values point-by-point).

(2) What is the *robustness* to alternative specifications of (b), (c), (d)? Which aspects of the problem formulation are critical, and which less so?

(3) Given the possibility of non-robustness, what is the potential for *data-dependent choices* of (b), (c), (d)?

(4) What is a good *measure of performance*? I share the author's disdain for error rates (Section 6). In particular, these global measures do not reveal any (spatial) pattern in the mis-matching between $\hat{x}$ and the known x^*. However in most real-life applications x^* is, of course, unknown. The important practical question is how to assess performance *in this case.*

(5) *Diagnostics*: Post-fitting, does the data like the model? It would be good to have graphical checks on the model (with or without formal tests).

(6) *Influence*: Are pixels at or near colour boundaries (collectively) highly influential on the reconstruction? Intuitively the answer seems to be yes. Given the (approximate) location of the colour boundaries, it only remains to allocate a colour to each of the regions they define.

(7) *Colour boundaries*: Discretisation of the scene into pixels implies that the measure of the colour boundaries is strictly positive. Now (6) suggests that pixels at or near a boundary together contain lots of information. Our suggestion is that we *look for them.* This idea may be of some value *per se* or as providing a starting point for *ICM*.

Dr F. P. Kelly (Statistical Laboratory, Cambridge): In the final section of this welcome and thought-provoking paper Professor Besag discusses the possibility of weaker restrictions on the conditional distribution $p_i(.\,|\,.)$. An application that springs to mind is the processing in real time of moving images. For example, consider the form

$$P(x_i(t+1)\,|\,y_i(t+1), \hat{x}_i(t), \hat{x}_{\partial i}(t)) \propto f(y_i(t+1)\,|\,x_i(t+1))p_i(x_i(t+1)\,|\,\hat{x}_i(t), \hat{x}_{\partial i}(t)),$$

and suppose $\hat{x}_i(t+1)$ is chosen to maximize this form with respect to $x_i(t+1)$. This might correspond to an approximation that $\hat{x}(t)$ is the true scene at time t, and a constraint that all that is known to the processor for pixel i at time $t+1$ is $y_i(t+1)$, $\hat{x}_i(t)$ and $\hat{x}_{\partial i}(t)$. The 'colour' $x_i(t)$ might now contain rate of change or edge velocity information. The important point is that a quite different set of considerations now affect the choice of the functional form $p_i(.\,|\,.)$.

A potential drawback of the *ICM* approach to the moving images problem is that tracking a time-varying local maximum may drag $\hat{x}(t)$ away from the time-varying conventional maximum likelihood classifier. This may be an argument in favour of a method which attempts to track the global maximum—for example by injecting additional noise into the step from $\hat{x}(t)$ to $\hat{x}(t+1)$ as in stochastic optimization algorithms.

Professor K. V. Mardia (University of Leeds): As others I found this paper very stimulating. I was first exposed to Image Processing problems in 1980 through some Landsat Data consisting of 3 rock types and 4 variables from Professor Switzer. Since the pixel size is over 1 acre, one expects pixels to be mixed; that is, at each pixel there may be a percentage of each rock type. Following on from the ingenious *ICM* method of Professor Besag, we developed a fuzzy classification method at Leeds University. My previous attempts were not successful to match the ground truth. Comparisons in 1 and 2 dimensions for various patterns show that one gets a smaller error of misclassification from a hardened fuzzy classification than from the *ICM* method. The parameter β measuring spatial correlation is fixed (at an optimal value) in both cases. The same remarks apply when one looks at the error on the boundary as well as at the interior points. Perhaps this is not surprising since for the fuzzy classification, the global maximum of the posterior density is the same as the local maximum. The fuzzy classification method is also worthy of serious consideration but, of course, the *ICM* method is faster and has other desirable properties.

In the different problem of target detection when group means etc. are unknown, we have found that the *ICM* method is sensitive to the starting value. A simple method of iterative local mean thresholding followed at each cycle by majority vote smoothing works very well in detecting the target object where the boundary pixels are not too small (see Mardia and Hainsworth, 1986).

It was disappointing to see that only single simulations were used in the illustrations for estimating the errors of misclassification etc. Further, the signals were only univariate. We have found that the optimal value of $\beta = 1.5$ for the univariate case becomes $\beta = 3.5$ for the 4-dimensional Landsat Data

with equal covariance matrices in the 3 groups. Intuitively, this increase in β with p is not surprising since the expected value of the quadratic form in the normal density increases with p.

Finally let me say again how much I enjoyed the paper.

Dr C. Jennison (University of Bath): I would like to point out the potential of Professor Besag's methods for producing a restored image of greater refinement than the original record. Fig. D4 shows a true scene divided into pixels, some of which contain areas of both of the two colours. A record was obtained by adding, independently for each pixel, Gaussian errors with mean zero and standard deviation 0.3 to the proportion of the pixel coloured black. Fig. D5 is the rather coarse reconstructed image produced by the *ICM* method using the model of Section 3 with $\beta = 1.5$. To refine this image we divide each pixel into four and allow each sub-pixel to take either colour in subsequent reconstructions. Only the aggregated record from each group of four sub-pixels is available but the *ICM* method can still be used to update groups of four sub-pixels at a time. With the model of Section 3 now applied to sub-pixels, this produces the 'refined' reconstruction shown in Fig. D6.

Since the original record was actually generated for each sub-pixel and then aggregated, we can examine the reconstruction that would have been obtained had the unaggregated record been available. This is shown in Fig. D7. There is little to choose between the reconstructions of Figs D6 and D7 but when

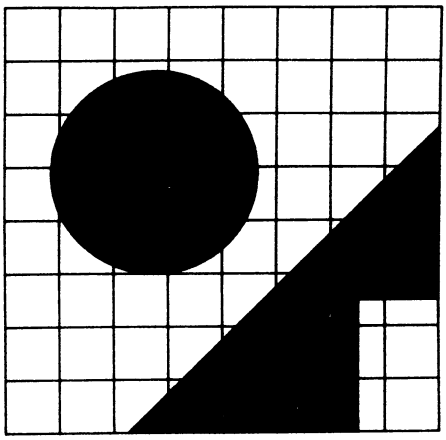

Fig. D4 True scene.

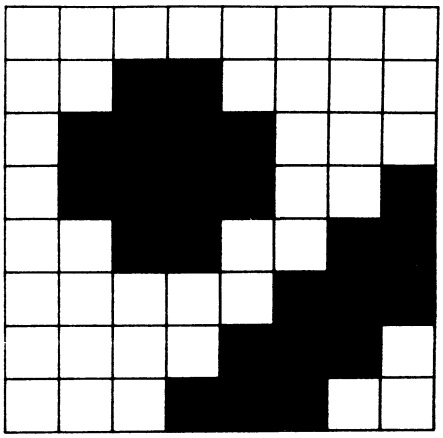

Fig. D5 *ICM* reconstruction on large pixel grid.

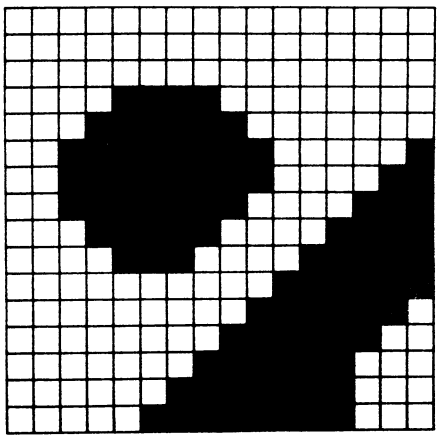

Fig. D6 Refined reconstruction using aggregated record.

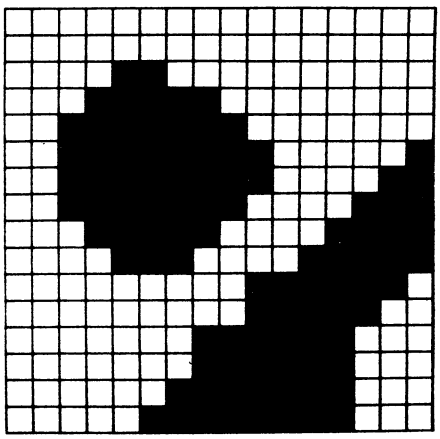

Fig. D7 *ICM* reconstruction from unaggregated record.

larger errors were added the reconstruction from the aggregated record was clearly superior. This surprising result could be a feature of working with 'blocks' of pixels in the initial stages of reconstruction, a step which is unavoidable when working with the aggregated record but which might be generally beneficial in its own right.

Dr J. T. Kent (University of Leeds): The stimulating and enjoyable paper we have heard tonight raises many issues about the statistical approach to image analysis. I will limit myself to one question about the general philosophy of modelling spatial dependence.

How important is it to have a realistic model of this dependence? For example in Time Series Analysis, a great deal of effort is often expended in determining a suitable *ARMA* model to fit the data. Thus a sophisticated appreciation of the whole correlation structure in the data is needed. On the other hand, in a field such as kernel density estimation, it is found that there is really only one important parameter, the window width, which affects the behaviour of the density estimate. The exact form of the kernel is not very important. Similarly, in tonight's paper, Professor Besag has tried to summarize spatial dependence in terms of a single parameter, β.

The limited work we have carried out at Leeds on spatial discrimination problems supports this simple approach. While the strength of spatial dependence is important to the accuracy of our algorithms, the exact form of spatial dependence is not so important. But perhaps this feature is due to the fact that no Markov random field model of low order is likely to be very realistic in spatial discrimination problems. Further, it may not be possible to answer this question in general; different problems in spatial analysis may require different levels of detailed knowledge of the correlation structure.

Dr D. J. Spiegelhalter (MRC Biostatistics Unit, Cambridge): A number of aspects in tonight's stimulating paper have parallels in current work on expert systems. Thus in place of the pixels in image analysis, we substitute possible clinical findings and diseases, while the regular lattice structure of the relationships expressed in the random field is replaced by a complex network comprising the local relationships archetypally expressed as subjective 'rules'. Whereas in image analysis a noisy realisation of all variables is available, for expert systems relatively precise observations are sequentially made on a subset, and from this fragmentary evidence the 'image' needs to be interpreted. Geman (1984) has, however, adapted the stochastic relaxation technique to this problem, although it is computationally expensive.

Alternatively, as a generalisation of the Markov mesh models mentioned in Section 1, it is much simpler if we can assume an initial ordering of the nodes in the graphical structure that expresses the 'knowledge'. The resulting directed graph corresponds to the 'causal network' representation popular in medical expert systems. (Pearl, 1986; Weiss *et al*, 1978). The work of Wermuth and Lauritzen (1983) and Kiiveri *et al* (1984) shows how an undirected graphical representation may then be obtained by re-expressing our beliefs as the particular class of 'triangulated' Markov random fields corresponding to a decomposable log-linear model. With this formulation, it is easy to compute exact marginal and joint posterior probabilities of currently unobserved features, conditional on those clinical findings currently available (Spiegelhalter, 1986a,b), the updating taking the form of 'propagating evidence' through the network.

In view of the above parallels with the stochastic relaxation and the Markov mesh models, it would be interesting to see if the techniques described tonight, which are of intermediate complexity, may have any applications in this new and exciting area.

Mr M. C. Shewry and Professor H. P. Wynn (The City University, London): We do not wish to miss an opportunity to congratulate Professor Besag on a very fine paper. There is a relationship to spatial sampling which follows on from the remarks of Dr. Critchley. If we are allowed to observe only at certain locations, in this case pixels, which pixels do we pick? A good criterion for optimum sampling design in this case is *maximum* entropy. To obtain the *minimum* expected posterior entropy we need to pick the sample with maximum entropy. For a stationary Gaussian process the optimum sampling is to choose the sites spread out roughly evenly over the region. However a model superimposed would tend to push observations to the outside of the region or to track the object in a moving image case. Our own work shows that as we move from the no model to the model situation holes start appearing in the region, where we are not allowed to sample. There is a close connection between the spatial sampling and the image processing models.

Dr Mark Berman (CSIRO, Sydney, Australia): I congratulate Professor Besag on this timely paper containing many interesting ideas. My own recent involvement in image processing has been with the restoration of degraded satellite and aircraft imagery, and for that reason I am particularly interested in Sections 4.1.3 and 4.2 of the paper. I wish to discuss further the model of Section 4.2.1 and restoration achieved via the m.a.p. estimator (15). First, it should be noted that the estimator as given is not location-invariant. This is easily corrected by subtracting the mean of X from both $\hat{x}$ and y in (15). However, for simplicity, I shall henceforth assume that X has zero mean.

It will be convenient to let $\Sigma_X (= Q^{-1})$ and Σ_Y denote the covariance matrices of X and Y respectively. Then (15) can be re-expressed as

$$\hat{x} = \Sigma_X \Sigma_Y^{-1} y. \tag{D1}$$

Although (D1) optimizes a variety of criteria, how similar are the properties of x and $\hat{x}$? Clearly, both have zero means. However, $\mathrm{cov}(\hat{x}) = \Sigma_X \Sigma_Y^{-1} \Sigma_X$ while $\mathrm{cov}(x) = \Sigma_X$. Hence, the second order spatial characteristics of x are not preserved by $\hat{x}$. To obtain an estimator, $\tilde{x}$, preserving second order properties, it appears necessary to diagonalize simultaneously Σ_X and Σ_Y. Specifically there exists an A such that $A\Sigma_X A^T = D = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ and $A\Sigma_Y A^T = I$. Because the signal and noise processes are uncorrelated, it follows that $0 \leqslant \lambda_i \leqslant 1$, $i = 1, \ldots, n$. One possible $\tilde{x}$ (satisfying certain optimality properties) is $\tilde{x} = A^{-1} D^{1/2} A y$; (D1) can be re-expressed as $\hat{x} = A^{-1} \mathrm{DA} y$. Since $0 \leqslant \lambda_i \leqslant 1$, $i = 1, \ldots, n$, we see that $\hat{x}$ is a heavier smoother of the data than $\tilde{x}$.

The potential problem with $\tilde{x}$ is the apparent necessity of orthogonalizing the data (or at least of obtaining A and D). This is not such a problem for processes where the signal and noise, as well as being uncorrelated with one another, are also both *stationary*. For then, the orthogonalization is approximately achieved via the Fourier transform. Restoration in the Fourier domain is a widely-used technique in image processing.

In light of the above, what are the implications for the m.a.p. and *ICM* estimators in Sections 4.1.3 and 4.2.2? Are there sensible modifications of these which are analogous to $\tilde{x}$? Are there any implications for the classification techniques described elsewhere in the paper?

Dr N. A. Campbell (Division of Mathematics and Statistics, CSIRO, Australia): The improvement in allocation performance quoted for the examples in the paper may not be realizable for multivariate data. Some results for four-band Landsat *MSS* data are as follows (from work done jointly with Dr Harri Kiiveri). Initial allocation rates of 97.0%, 79.5% and 76.6% for three classes (overall rate of 80.7%) from the usual allocation procedure are improved to 95.3%, 96.8% and 91.3% (overall rate of 94.4%) after 5 iterations; after 2 iterations, the values are 95.3%, 96.3% and 88.0% (overall rate of 92.9%). A value of $\beta = 1.0$ with a second-order neighbourhood is used, the usual allocation map providing the initial labels. A value of $\beta = 2.0$ gives allocation rates of 89.9%, 97.6% and 90.5% (overall rate of 93.8%), indicating the relative lack of sensitivity of the results to the choice of β (see Section 3).

Introducing a conditional autoregressive (*CAR*) model for the records, together with a prior label model, and starting with the labels from the above solution, gives a slight improvement in allocation performance, to 96.5%, 96.9% and 92.3% (overall rate of 95.0%) after 2 iterations. Starting from the usual allocation labels gives a marginally worse allocation performance than the label model alone (by around 1–2%). The importance of the initial labels for an improvement in allocation performance needs to be stressed. A *CAR* component can lead to poorer allocation performance unless there is a good agreement between the initial labels and the actual labels. Otherwise, the neighbour-correction and standardization which arises from the *CAR* model is inappropriate.

It is interesting to note the values of the posterior probabilities associated with (5) as the iterations proceed. For β around 0.75 to 2.0, the posterior probabilities for pixels interior to a class tend rapidly to 0 or 1; the label model in (6) acts as a strong local prior which dominates the calculation unless the signal for one of the other classes is exceptionally strong. It is only on class boundaries, where the neighbour information may be more equivocal, that the posterior probabilities sometimes lead to a doubtful allocation.

Dr Rodney Coleman (Imperial College, London): I welcome this paper which shows how statisticians can contribute to the developments in image analysis.

The use of local weighted-average smoothing filters has long been the basic practice in image analysis (Russ and Russ, 1984). The novelty here is their application to unordered colours, their use iteratively, and the statistical theory. In particular, in computer tomography, the processed images show clearly

defined and recognizable shapes. This is the result of local smoothing built into the machines by its engineers. In the absence of any theory this leads me to believe that almost any local smoothing would be effective. The numerical inversion of the formulae in Wicksell's corpuscle problem gives disappointing results when the data are grouped and we discretize the integral expressions, whether or not we have noisy data (Blödner *et al*, 1984). Yet this is a close relative of the formal procedures in analyzing tomographic data. The difference lies in the use with the Wicksell problem of global constraints. Local smoothing using kernel estimators has been used in the Wicksell problem, and the similarities between this and the *ICM* method of this paper have already been pointed out by Dr Kent.

It is worth noting that image enhancement is not always to be equated with a correction for noise. We may wish to remove true local variability: for example, to reduce 32 true colours to 4; or in a soil or agricultural survey to give a single colour to each county.

Dr I Ford and Mr. J. H. McColl (University of Glasgow): Professor Besag alludes to the need for image-processing in nuclear medicine. We are particularly interested in statistical aspects of image reconstruction and analysis in positron emission tomography (*PET*), single photon emission tomography (*SPECT*) and the related area of autoradiography. (Phelps, Mazziota and Schelbert (1986) is a useful general introduction to the areas of PET scanning and autoradiography.)

These techniques produce information about pixel colour in the form of Poisson counts. Poisson data are obtained directly in autoradiography, where the Poisson counts may be assumed independent between pixels. This would provide a source of data with non-Gaussian noise (see Dr Kay's comments). In *PET* and *SPECT* the observed Poisson counts correspond to line integrals. This means that the estimated pixel counts are not independent, since any reconstruction algorithm inevitably introduces some form of local smoothing.

In these medical applications there is not in general any great interest in the aesthetic properties of the global image itself, the image merely being a by-product in the process of obtaining information about some underlying functional parameter, such as bloodflow, in regions of the body of interest in the current investigation. We believe that smoothing of the global image might be dangerous when interest lies only in local areas.

This problem is compounded if, as might occur in cerebral *SPECT* imaging, the underlying parameter varies systematically from brain region to brain region and the brain regions of interest have areas of the same order of magnitude as the pixels. For example, a relatively small black region embedded in a relatively large white region might be reconstructed as light grey. Reconstruction errors of this type, known as 'partial voluming', are well known in *PET* scanning.

We would welcome comments from Professor Besag on our suggestion that his algorithm, while possibly improving global properties of an image, might cause a deterioration in important local properties.

Finally, we ask how we might assess the reliability of any reconstructed image. As statisticians we are generally not happy with merely calculating a point estimate for a parameter. Is there any method of determining how accurately our estimated image represents the true underlying scene?

We welcome Professor Besag's paper as an important initial step in a developing area of statistical application.

Professor Donald Geman (University of Massachusetts, Amherst): It is a pleasure to commend Professor Besag for his insightful investigation of several important issues in Bayesian image reconstruction: (1) The utility of Markov random fields as image models, and the distinction between their small- and large-scale properties; (2) The related problem of formulating suitable performance criteria and the corresponding question of local versus global reconstruction; and (3) The estimation of model parameters in the face of intractable means and information loss.

There are numerous connections between this paper and one currently in preparation by Stuart Geman, Donald E. McClure, and myself. We have also tried to estimate parameters in exponential families in very high dimensions with incomplete data. One may observe only a component of the model (e.g. grey levels), the others, such as edges or labels, being hidden; or the degradation may be of the type here, e.g. additive noise.

Concerning 1) and 2), and the use of *context*, one can envision a variety of algorithms depending on one's evaluation of loss, computational feasibility and mathematical coherence. The role of spatial information may depend on the selection of a loss function as well as the prior model. Thus, for instance, with a zero-one loss function, one is led to the m.a.p. estimator and hence to consider *all* of the labels

simultaneously given all of the data, whereas with performance measured by the percent of misclassified pixels one seeks the univariate posterior modes. Other algorithms are not driven by minimizing penalties. Professor Besag uses the maximum likelihood classifier (maximizing each pixel label given only the record there) as the starting point for *ICM*, in which one successively maximizes each label given the neighbouring labels and all the data.

ICM is a form of iterative improvement on the posterior likelihood. We have also used (local) "quenching" in experiments on segmentation and tomography, and one often arrives at a "satisfying" local maximum. In boundary-finding, our model accounts for intensities, micro-edges, and boundaries (macro-edges). The posterior is the conditional distribution on edges and boundaries given the (undergraded) intensities. We use stochastic relaxation to find a starting point and then a form of iterative improvement referred to here as "block reconstruction" to obtain a final labeling. In the case of texture segmentation we have stuck with annealing.

Obviously these issues are wide-open and ripe for further study. Professor Besag has made a most sensible start.

Professors S. Geman and D. E. McClure (Brown University, U.S.A.): Professor Besag's paper presents a timely and cogent discussion of alternative loss functions that can be used to determine algorithms for image reconstruction based on Bayesian models. The use of probabilistic models in image analysis has opened new possibilities for basing algorithms on a firm scientific foundation, and the clear connection of the methodology to the principle of minimum Bayes risk is pivotal for this foundation.

We have used the *ICM* algorithm for recent computational experiments in single photon emission computed tomography (*SPECT*; section 4.2.2) and attest to its effectiveness when the iterative reconstruction is started with a good initialization. We firmly believe that *special purpose* algorithms of this type—a judicious initialization together with simple suboptimal iterative improvement—are very important for the utility of the Bayesian methods in actual applications. By adapting the algorithm to the application, one can take advantage of special structure in the underlying model or of good alternative reconstructions to reduce the computational burden.

The *SPECT* example is an interesting departure from Professor Besag's Assumption 1 and the example in Section 5.2, where observables y have a local dependence on the true image x. In *SPECT* the dependence is manifestly global, and hence the posterior distribution $\pi(x|y)$ has nonlocal neighborhoods. Nonlocal neighbourhoods increase the computational burden, but they do not contradict the conceptual foundation of the model-based reconstructions.

While special purpose algorithms will determine the utility of the Bayesian methods, the general purpose methods—stochastic relaxation and simulation of solutions of the Langevin equation (Grenander, 1983; Geman and Geman, 1984; Gidas, 1985a; Geman and Hwang, 1986)—have proven enormously convenient and versatile. We are able to apply a single computer program to every new problem by merely changing the subroutine that computes the energy function in the Gibbs representation of the posterior distribution.

Mr J. Haslett (Trinity College, Dublin): There are, by now, a number of methods for restoring noise corrupted images. Many of these will achieve similar restorations, and not only as measured by the per pixel error rate. The emphasis by now is on the second order differences between these methods.

Professor Besag has concentrated on the differences between two 'statistical' methods, his own and that of Geman and Geman. The practical differences are computational, although there are some interesting insights into the use of Markov random fields. Both achieve very similar reconstructions to the 'non-statistical' method of relaxation, popular for many years in image processing.

The second order questions to which statisticians might be expected to contribute include the questions that are normal in any statistical classification procedure—accuracy overall (error rate in a given study) and posterior probabilities for a given case (pixel)—as well as the identification of outlying cases (pixels).

Indeed overall accuracy is the first question to be posed by any potential client thinking of using remote sensed imagery as opposed to traditional ground based surveys in, for example, an ecological inventory. At a higher level in image analysis, in terms of identifying objects—such as a cancerous growth—the same sort of questions arise: what is the posterior distribution of the size of the tumour which is suggested by image analysis?

It seems from Professor Besag's paper that *all* such questions can be answered (in principle) by sufficient simulations at a temperature $T = 1$, under Geman and Geman's method, but that *no* such questions can be derived from the *ICM* method. This latter derives from the notorious difficulty, already

referred to, of deducing any marginal distributions from the Markov random field model. It seems, therefore, that we are being offered no alternative, as statisticians, to the route of vast raw computing power being pioneered by the Gemans.

Yet simple solutions, though imperfect, do exist to some of the easier problems above. The Markov mesh models, and the Pickard model do yield explicit posterior probabilities at each pixel. A linear method (Haslett and Horgan, 1985) has achieved, in addition, a fair degree of accuracy in predicting the overall error rate in binary images corrupted by noise and very recent work indicates that this can be extended to the case of many underlying classes, and in addition, can assist in identifying outliers. The key to all these is the abandonment of the Markov random field model, a step which Professor Besag, it appears, has already taken himself!

Dr J. W. Kay (University of Glasgow): May I congratulate Professor Besag on his most stimulating paper. The *ICM* algorithm may be viewed as a decision-directed relaxation labelling algorithm. Faugeras and Berthod (1981) provide the following 'compatibility function' to represent the local spatial relationship between the true colours of the pixels in the neighbourhood of the ith pixel:

$$p_i(x_i|x_{\partial i}) = \sum_{\beta_{\partial i}} p_i(x_i|x_{\partial i} = \beta_{\partial i}) \prod_{j \in \partial i} p_j(x_j = \beta_j)$$

where $\beta_{\partial i} = \{\beta_j\}$ is a specific assignment of labels to $x_{\partial i}$.
Now, if x_j is replaced by it's current estimate $\hat{x}_j$, then the right-hand side of this equation becomes

$$p_i(x_i|\hat{x}_{\partial i})$$

which appears in equation (5) of the paper. Hence decision-directed relaxation labelling methods which employ the feature data during reconstruction (Kay and Titterington, 1986), are equivalent to the *ICM* method.

Relaxation Labelling methods generally exhibit a quick initial improvement, especially when acceleration techniques are used (Zucker *et al.*, 1978) and involve local iterations; so does the *ICM* algorithm. Kalayeh and Landgrebe (1984) provides an adaptive relaxation labelling algorithm which is similar in spirit to the *ICM* method, both using a local estimate of the image to construct $p_i(x_i|\hat{x}_{\partial i})$. However, decision-directed methods are prone to inconsistency and runaway (Young and Calvert, 1974); also, Richards *et al.* (1981) show that the relaxation labelling process can suffer from degradation in the sense that image features e.g. corners, edges are lost during the reconstruction process. Can this also be a problem for the *ICM* algorithm?

Would the author please comment on the performance of the *ICM* method (a) under different noise regimes and (b) under different noise-signal interactions? Does *ICM* work as effectively as in the additive Gaussian noise case? What values of β would be appropriate?

Finally, some general comments. It should be noted that many real image-analysis tasks involve sequences of images e.g. angio-cardiography, foetal ultrasound scans, so that a temporal dimension is present. Also, it is important that the task of image reconstruction is *not* viewed as an end in itself, but rather as a necessary process before image analysis proceeds. Hence often reconstruction methods will be required to serve the needs of later image analysis and so they will need to be *feature-preserving*. Would the author care to comment on how the *ICM* algorithm may be extended to contribute towards these more complex tasks?

Dr H. T. Kiiveri (CSIRO, Wembley, W. Australia): I would like to comment on two points. The first concerns the *ICM* algorithm and the second the algorithm given in Section 5.1.2.

It is not clear that the *ICM* method of reconstruction ignores the large scale deficiencies of $p(x)$. It would seem that both global and local maxima of $P(x|y)$ are determined by the form of $l(y|x)$ and $p(x)$. Since *ICM* is simply a cyclic ascent algorithm, the choice of initial scene may be quite crucial to the success of the reconstruction. If the initial scene is sufficiently far away from the true scene, the algorithm may converge to an undesirable local maximum, one which is determined by the large scale deficiencies of $p(x)$.

The algorithm in Section 5.1.2 is an approximation to a general algorithm, the approximation being necessary because of the intractable normalising constant in $p(x)$. The general algorithm is defined below, where L denotes the joint distribution of x and y.

(1) Given x_n, θ_n, ϕ_n choose θ_{n+1}, ϕ_{n+1} such that $L(x_n y|\theta_{n+1}\phi_{n+1}) \geqslant L(x_n y|\theta_n \phi_n)$ e.g. maximise $l(y|x_n\theta)p(x_n|\phi)$ to get θ_{n+1}, ϕ_{n+1}.

(2) Choose x_{n+1} so that $L(x_{n+1}y\,|\,\theta_{n+1}\phi_{n+1}) \geqslant L(x_n\,|\,\theta_{n+1}\phi_{n+1})$ i.e. choose x_{n+1} so that $P(x_{n+1}\,|\,y\theta_{n+1}\phi_{n+1}) \geqslant P(x_n\,|\,y\theta_{n+1}\phi_{n+1})$.

This algorithm has the property that $L(x_{n+1}y|\theta_{n+1}\phi_{n+1}) \geqslant L(x_n y|\theta_n\phi_n)$. The monotonicity of L and the fact that x_n is contained in a compact set suggests that the procedure (with an appropriate stopping rule) will converge. However a rigorous proof would need to consider the existence of θ_n, ϕ_n. The algorithm is similar to an *EM* algorithm with step 2 replacing the intractable *E* step. Note that the algorithm in Section 5.1.2 corresponds to using a pseudolikelihood instead of $p(x_n|\phi)$ in step 1 and one cycle of the *ICM* or cyclic ascent algorithm for step 2.

A second algorithm can also be suggested. This algorithm is perhaps more in the spirit of a Bayesian approach to the reconstruction problem. Given x_n, θ_n, ϕ_n

(1) Choose θ_{n+1}, ϕ_{n+1} so that $P(x_n|y\theta_{n+1}\phi_{n+1}) \geqslant P(x_n|y\theta_n\phi_n)$

(2) Choose x_{n+1} so that $P(x_{n+1}\,|\,y\theta_{n+1}\phi_{n+1}) \geqslant P(x_n\,|\,y\theta_{n+1}\phi_{n+1})$.

It easily follows that $P(x_{n+1}|y\theta_{n+1}\phi_{n+1}) \geqslant P(x_n|y\theta_n\phi_n)$. The algorithm can be made a practical proposition by using a pseudolikelihood (posterior) in (1). The second step can be performed as usual.

Professor Hans R. Künsch (ETH Zurich): I am very much impressed by Professor Besag's interesting paper, in particular because his simple method seems to work so well. In image reconstruction we meet again the basic problem of smoothing: How can one remove all random irregularities without destroying parts of the finer structure of the data? Professor Besag's method pays more attention to the second problem as Fig. 1c shows: the peninsula towards the bottom is retained, but the reconstruction still contains some random effects like the small cluster of dots in the lower left part. I would be interested very much to see how some of the methods mentioned in Sections 2.3 and 2.4 work on the same data sets. In particular, maximizing $P(x_i\,|\,y_{\lambda i})$ is optimal among all reconstructions $\hat{x}_i$ depending only on $y_{\lambda i}$ if we choose the error rate as our criterion. *ICM* is computationally much simpler, but how much does it loose?

Finally, let me make a remark about parameter estimation. Maximum likelihood is intractable for the models considered by Professor Besag, but a one step approximation starting from one of the simple alternatives might be feasible. For simplicity consider the estimation of β in the model (6) of Besag's paper, other cases being similar. The one step approximation is then

$$\hat{\beta}^{(1)} = \hat{\beta} + \mathrm{var}_{\hat{\beta}}[V]^{-1}(V - E_{\hat{\beta}}[V])$$

where $\hat{\beta}$ is the original estimator and V denotes the number of neighbour pairs having like colours. The expectation and variance of V have to be found by Monte Carlo. This is still computationally expensive and thus not suitable for routine purposes, but in certain situations it will be useful. Note that in order to estimate the variance of $\hat{\beta}$ similar calculations are needed anyhow. The above one step approximation has the theoretical advantage of being asymptotically efficient as the number of pixels increases. Details will be given elsewhere.

Dr Antonio Possolo (University of Washington): I have had the opportunity to experiment with the full-fledged version of *ICM* (as in 5.1.2, where both θ and ϕ are estimated concurrently with the image), and it gives me great pleasure to second Professor Besag's claim that it works. The model-tempered approaches to image reconstruction favoured by both Professors Besag and Geman are most welcome alternatives to the bric-a-brac of *ad hoc* techniques that superabound in the field. Future trials may well establish that the products sponsored by Professor Besag are the best all-purpose image cleaners ever.

In one experiment, the image $\{x_i\}$ was taken as an outcome of a binary m.r.f. $\{X_i\}$ on a 100×100 grid, distributed according to Verhagen's (1977) model, as described by Pickard (1977), with $p(x) \propto \exp\{\phi' U(x)\}$, where ϕ is five-dimensional. The observations $\{y_i\}$ were obtained by *flipping* each x_i to $1 - x_i$ independently, with probability $\theta = E(X_i)$. In this particularly adverse situation, employing m.l.e. for θ and my logit procedure (Possolo, 1986) for ϕ, on the fly, in each cycle of *ICM*, typically yielded reconstructions of quality comparable to the best reported by Professor Besag.

It seems best to initialize $\{\hat{x}_i\}$ as outcomes of random elements with appropriate marginal distributions (*cf.* setting the temperature high at the beginning of annealing). In the images I have experimented with, the components of ϕ were largish in absolute value, and the estimation of ϕ during *ICM* emulates simulated annealing in that $\hat{\phi}$ evolves in a direction that corresponds to lowering the value of T in (4).

It is a 'data driven' method of accomplishing what Professor Besag acknowledges works best: 'β increasing ... resulted in an improvement'. Still, one should welcome further clarification on the relationship between *ICM* and m.a.p. reconstruction via simulated annealing. In particular, $\hat{\phi}$ seems to converge towards ϕ at a progressively slowing pace, and ends up moving at a crawl (say, one part in one thousand per iteration) around $c\phi$, for some $0 < c < 1$—this resembles stopping annealing too early.

In Section 3, Professor Besag suggests that setting the (one-dimensional) parameter β to 1.5 'seems to work well ... though the precise value is not crucial'. When ϕ is multidimensional, it is my experience that its value matters, and one is well advised to estimate it during *ICM*. The way it matters, however, is consistent with what Professor Besag has found to be adequate when $\phi(\equiv \beta)$ is one-dimensional: the reconstruction procedure is robust in the sense that a value of $\hat{\phi}$ of the form $c\phi$, for quite a range of positive c's not necessarily close to 1, will yield acceptable accurate reconstructions.

Dr E. M. Scott (University of Glasgow): I would like to congratulate Professor Besag on a very interesting paper which has dealt with the processing of images. I would now like to consider briefly the analysis of the processed images in a particular application.

My interest in image processing arises in the following context: remote sensing of environmental parameters such as land use and forest resources and their changes as a result of agriculture and forest clearing. These parameters are of particular interest as input into models of the global carbon cycle. Man's intervention in the natural biospheric processes provides a major perturbation of the cycle with far reaching effects such as changes in the climate e.g. the greenhouse effect.

Satellite imaging of land use and deforestation provides one means of quantifying the changes in these parameters, and so within the reconstruction procedure there must be some method of comparing sequential frames. This problem might also arise in the ultrasonic imaging of foetuses where between two frames there may well have been considerable development of the foetus. So my first question is how one might detect significant changes between reconstructions and in a related manner, how one might measure the uncertainty on attributes identified in the reconstruction?

A more general point would concern the use of a global criterion such as error rate to select parameter values and to assess reconstructions. In particular applications, other, perhaps more local, criteria could be of more practical significance. I would suggest that measures of shape, for instance, might be of particular importance. My second question then is can the *ICM* algorithm be modified by constraining the procedure to be sensitive to the existence of boundaries of attributes known to be in the true image?

Professor P. Switzer There are just a few points I would like to raise in connection with Julian Besag's stimulating paper. Although it certainly was not the intention, there is an impression left that one could obtain very low classification error rates even in complicated situations if a clever algorithm were used. But the simulated examples which illustrate this point may be atypical in that their real strength comes from using spatially uncorrelated noise in combination with single-colour groupings that seem to consist of ten or more pixels. With this setup classification errors with naive procedures seem to occur ubiquitously at random while the few errors of a contextual classifier occur along boundaries between classes. However, with real images, classification errors often occur in patches. The usual reason for this is that the patch is atypical, with respect to the measured values, of the class to which it belongs. This may be due to illumination effects, textural properties, dying back of vegetation or what have you. To model this common situation properly one would at least wish to use randomness with significant spatial autocorrelation within classes, even without the complication of possibly multimodal frequency distributions within classes.

Contextual classifiers which make use of neighbour pixel information are expected to do well in the interior of large contiguous areas of a single colour. Serious competitors should then be compared on their performance near class boundaries. Then it appears one wants a classifier which uses neighbour information in a way which distinguishes between probable and improbable class configurations within neighbourhoods. For example, a 'majority-vote' updating algorithm seems to be indifferent to the actual pattern of class assignments within neighbourhoods. This raises the question of whether the adopted model of the underlying class pattern itself generates reasonable behaviour near the boundary between classes. Indeed, it seems reasonable that the boundary between classes should not be made to follow pixel boundaries. In such cases no pixel crossed by a class boundary could be correctly classified and there would be a non-trivial lower bound for the classification error rate. Where the size of contiguous homogeneous areas is typically on the order of a few pixels then boundary questions come to dominate; the problem may then change to one of estimating class fractions for each pixel.

Dr B. Torsney (University of Glasgow): Professor Besag's paper is of particular note in that it extends the application of some existing statistical techniques to a further family of problems. This is also true of techniques developed in the optimal design arena.

When ambiguous labelling is allowed we conceive that to each pixel there can be associated a probability distribution on the colours. An optimization approach such as that of Faugeras and Berthod (1981) seeks a set of such distributions to optimise some criterion depending say on ambiguity and consistency. This defines a generalisation of the following problem: (P1) maximise $\phi(p_1, p_2, \ldots, p_J)$, subject to (i) $p_j \geqslant 0$, (ii) $\Sigma p_j = 1$.

Several examples of (P1) arise in the statistical domain, particularly the optimal linear regression design problem. It was with this case in mind that Torsney (1977, 1983) and Silvey *et al.* (1978) proposed the iteration $p_j^{(r+1)} \propto p_j^{(r)} f(d_j)$, where $d_j = \partial\phi/\partial p_j^{(r)}$ and $f(d)$ was taken to be d^δ, with δ being a free positive parameter.

Of primary relevance here is that the principle of this iteration seems identical to that of relaxation labelling techniques. See Rosenfeld *et al.* (1976), Zucker *et al.* (1978, 1981), Hummel and Zucker (1983), Boyce *et al.* (1985). We note also that in employing a free parameter there is a similarity with the author's *ICM* method.

One particular parallel development emerges below. The image reconstruction problems, having several probability vectors, are special cases of a problem which is obtained if the constraint $\Sigma p_j = 1$ is replaced by several linear equality constraints. This problem was considered by Torsney (1981) and in an unpublished contribution to the society's conference at the University of York in 1982. In the latter case the problem of finding the maximum likelihood estimates of the cell probabilities of a 3×3 contingency table under the hypothesis of marginal homogeneity was transformed to an example of (P1), and the above iteration with $f(d) = \exp(\delta d)$ was explored for various values of δ. The case $\delta = 1$ yields the algorithm of Boyce *et al.* (1985).

It is intended to futher explore these links.

I am indebted to my colleague Jim Kay for introducing me to the relaxation labelling literature.

I would like to conclude with the warning that Professor Besag's title is open to misinterpretation. All good thinking members of the society will of course take it for granted that he wishes to clean up dirty pictures, but a more literal interpretation of the title leaves him open to the accusation that he simply wishes to analyse, albeit statistically, if not exploratorally, the contents of such images!

The Author replied later, in writing, as follows.

I am extremely grateful to the discussants for their valuable comments and criticisms, which have given me much to think about. Indeed, I have found it extremely difficult to compose a detailed reply: the more I have written, the more vague it has seemed to become and the more deadlines have passed me by! I think the main problem is that spatial statistics in image processing is still short of a real success story, despite the large number of potential applications: perhaps the Brown/UMass group is closest, with their work on tomography, for example. Although the lack of success is rather to be expected, given the relatively small involvement by statisticians, thus far, and the almost essential need for specialized equipment, nevertheless it means that one's comments are necessarily speculative and this eventually becomes unsatisfactory. However, I shall do my best to address some of the points raised in the Discussion and hope that my numerous sins of omission will be forgiven: also some of the most valuable contributions require little or no further comment.

I have organized my reply under a number of headings but I should perhaps emphasize that the space devoted to any particular topic is not intended to indicate its relative importance! Despite Professor Titterington's remarks, I have become increasingly enamoured with the Bayesian paradigm, in the present context, and will adopt the associated terminology, though this is by no means essential and somewhat in conflict with the pragmatism, both in *ICM* and in parameter estimation.

The likelihood

Probability has its most obvious role in linking the records y_i to the original image x through a specified probability density or likelihood $l(y|x)$, perhaps also depending on a few unknown parameter values. More stringent demands, such as conditional independence (cf. Assumption 1 in Section 2.1), simplify computation but are not essential; and the records may provide only an indirect rather than a direct representation of the true scene, so that applications such as computerised tomography are included (Geman and McClure, 1985). In the context of a particular practical problem, the central issue is more fundamental: is any simple probabilistic description of the records useful or even appropriate?

Thus, in nuclear medicine, the answer seems clear, for it follows from physical considerations that a Poisson formulation is indeed correct; furthermore, the associated variability is often a major cause of difficulty in reconstructing a faithful estimate of the true image, whether or not this is compounded by problems of attenuation and tomography as in the discussion by Professors Geman and McClure. On the other hand, the position is perhaps less satisfactory in the processing of satellite data, such as those available from Landsat. There, it is sometimes assumed that, given the ground-truth classification x, the multi-component y_i's follow, possibly dependent, multivariate Normal distributions, though it is not clear, to me at least, that this or any similar stochastic formulation successfully describes the principal ambiguities in the data. The warnings of Professor Switzer are particularly apposite here and suggest that the findings of Dr Campbell and Dr Kershaw are not entirely surprising. One possible alternative, still involving spatial considerations, would be to segment the image into apparently homogeneous regions, each of which could subsequently, or concurrently, be given a classification; but entirely different techniques may hold the real key to satellite images.

The prior

A second role for probability occurs in ascribing the characteristics of the true scene to a stochastic process $\{p(x)\}$ or, in the Bayesian framework adopted here, in reducing one's prior beliefs about the true scene to a probability distribution. Several discussants comment on the choice of prior and on its possible subsequent effects. On both counts, I rather regret having used only the simple distribution (6) in illustrating *ICM* and some of its variations, for this could easily create the impression that (6) is the "standard" prior for unordered colours, perhaps with minor modifications to incorporate distance- and direction-dependent β's. However, (6) was intended merely as a plausible representation of "prior ignorance" about the true scene. Nor should the value $\beta = 1.5$ be taken too seriously: certainly there has never been any claim of "optimality". Indeed, it was suggested to me long ago that $\beta = 1.5$ is really too large. I stuck with it in the paper because (i) it seemed good enough for illustrative purposes on artificial data and (ii) the exact value of β is usually unimportant with *ICM* if, as recommended in Section 2.6, smaller values are used on earlier iterations. An extreme instance of the second point occurs in a slightly different context in Fig. 6c, where the final pseudo-likelihood estimate of β is infinite, resulting in a majority-vote pattern. Of course, to use majority-vote *throughout ICM* or to use too large a value of β with m.a.p. can easily have disastrous consequences (cf. the contribution from Drs Greig, Porteous and Seheult). A final point about β in (6) concerns Professor Mardia's comments on multivariate records. There can be no formal justification in varying β according to the number of channels; and a value as large as 3.5 stretches probability considerations to implausible extremes! Of course the effect of any particular prior will diminish as the number of channels increases, provided the additional channels are informative, and this is as it should be. Incidentially, the original version of the paper included a four-channel example, with results entirely in accord with those for univariate records.

The major issue is whether we can get away from such simplistic distributions as (6) to produce priors that successfully capture known stochastic components of the true scene. Dr Green shows how this can be achieved for global features, using his modified prior $\{p^*(x)\}$: in particular, this allows the approximate maintenance of a specified marginal frequency distribution, whether for unordered colours or for grey-level scenes. In practice, such frequencies might be chosen in accord with the maximum-likelihood classifier, close in spirit to empirical Bayes estimation. As regards local characteristics, some broad suggestions appear in Section 4 of the paper, from among the class of pairwise-interaction distributions (9). One key notion is the availability of the conditional distributions $\{p_i(x_i;\, x_{\partial i})\}$, an aspect which is particularly attractive for unordered colours. Note also that when relevant training data are available, the conditional distributions can be estimated from their pseudo-likelihood or otherwise.

A more sophisticated and adventurous formulation of the prior is through "metric pattern theory" (Grenander, 1983). Here the aim is to attach a probability distribution, not only to the pixel colours or grey levels, but also to supplementary features of the scene that may not be directly observed. For example, Geman and Geman (1984) use "line sites" which sit on the edges of pixels and whose states indicate whether any given edge is part of a boundary between two colours and, if it is, possibly also its orientation. The part of the prior dealing with line sites encourages continuation of the boundaries. The potential improvement in restoration, over a prior such as (6), is illustrated in Fig. 4 of Geman and Geman (1984). Much more intricate priors for boundary detection have been devised by Professor D. Geman, as he mentions in his discussion. Similar notions might be useful with satellite data, perhaps using texture models within regions. Another possibility is to model gradients in grey-level scenes.

My own rather limited use of metric pattern theory concerns support for locally linear features in the

image. For example, a terrestrial scene may contain roads, perhaps no more than one or two pixels wide and easily destroyed by conventional smoothers or a (clearly inappropriate) prior such as (6). One possible remedy is to examine the difference between smoothed and unsmoothed images but a much neater solution is to augment the colour "road" by various markers, representing straight and curved segments, junctions, cross-roads and dead-ends, each with its orientation. The prior should encourage desirable combinations on adjacent pixels, so that the roads tend to form a continuous network. Note, first, that in general the maximum-likelihood classifier no longer provides a unique description of the scene, including markers, and cannot be used to initiate *ICM*: see Professor D. Geman's comment. Second, the models of metric pattern theory need only involve pairwise interactions but now on the extended state-space: my remark in Section 4.3 may have been misleading here. Finally, the very existence of markers implies a significant increase in the computational burden. Thus, it may be prudent to consider less elegant models, based only on pixel colours but involving terms for third- and fourth-order cliques (cf. Dr Kittler's remarks), which can therefore take account of the actual pattern of colour assignments within neighbourhoods (cf. Professor Switzer's comment). Clearly, there is much scope for further research but, in the meantime, I hope that models such as (10) and their continuous analogues will find their uses.

Several discussants express concern about the effects of the prior on the restored image. The first danger, as mentioned by Dr Kay and Dr Ford and Mr McColl, is that small but genuine feature will be smoothed out of the scene. This may of course happen but at least there are safeguards. First, the original records are retained at every stage of processing. Second, the prior can be chosen to accommodate anticipated awkward features, as discussed above. Third, restoration should depend primarily on the local characteristics of the prior, unless special global modifications have been made. Finally, the difference between the restored image and a simple context-free estimate, such as the maximum-likelihood classifier, can be examined for anomalies. The opposite danger is that the prior may introduce spurious features into the scene. This seems to me to be a lesser risk. It should be noted that the realizations of M.r.f.'s which appear in the literature are often highly atypical and arise because the space-time simulation has been prematurely curtailed. For example, the prior in Geman and McClure (1985) has realizations which are typically rather flat and not as in their Fig. 2A, though, by design and in effect, the prior does indeed cater for sudden discontinuities suggested by the records. A third danger is that as the prior becomes more intricate, so it becomes more difficult to interpret, both by the user and by his or her algorithm.

Choice of method

In Section 2 of the paper, I discussed three methods of estimating the true scene, each derived from the posterior distribution $\{P(x|y)\}$. In summary, these were: (a) m.a.p., in which the estimate $\hat{x}$ is chosen to maximize $P(x|y)$, corresponding to a 0–1 loss function on the true scene; (b) the marginal posterior modes, which maximize $P(x_i|y)$ for each i and therefore minimize the posterior expected error rate; and (c) *ICM*, an iterative procedure based on conditional modes. Some interesting connections with other techniques are mentioned in the Discussion and it is of course prudent to take advantage of these, whenever possible. However, as a statistician, I want to make use of probabilistic reasoning, whenever I can. Thus, in response to Professor Titterington, I *want* the prior to be one that I believe in, or at least one whose salient characteristics for restoration I believe in; and I also strongly favour a procedure which copes automatically with Poisson records, say. As regards a choice between (a), (b) and (c), though (c) only partially fulfils the above requirements, I make a few additional comments below.

As regards (a), the deterministic algorithm of Drs Greig, Porteous and Seheult is an important contribution and demonstrates the susceptibility of m.a.p. to the large-scale characteristics of the prior. Note also that simulated annealing, on a fixed "temperature" schedule, produces increasingly poor approximations to m.a.p., as the interaction parameter β increases. One might expect this type of result to hold rather generally, as the stochastic algorithm gets increasingly bogged down by local maxima. Indeed, I contend that, because of this, simulated annealing will usually produce more acceptable results than m.a.p., unless the prior has been modified as in Dr Green's contribution. The deterministic m.a.p. algorithm allows such conjectures to be properly investigated, at least for binary patterns. However, even in an ideal setting, m.a.p. has drawbacks: (i) it is generally sub-optimal for any non-linear functional of X, as remarked by Professor Silverman, and (ii) confidence statements are unavailable, despite the vast computational effort, as mentioned by Mr Haslett. Of course, *ICM* also suffers from (ii) and, in general, has no optimality properties. Furthermore, as Dr Kiiveri stresses, the algorithm depends critically on the initial estimate, since it cannot escape from any local maximum of $P(x|y)$. On the other hand, it seems reasonable to suppose that the maximum-likelihood classifier, suggested as an integral part of

ICM in the paper, will generally provide a reasonable first approximation to the true scene and that the iterative process will then smooth successive estimates in accordance with the local characteristics of the prior. Under this regime, I surmise that *ICM* will usually outperform m.a.p., again unless careful global modifications have been made to the prior. And the speed with which *ICM* typically converges means that it is a contender for real-time or limited-resource image processing.

As regards (b), I should like to return briefly to the final paragraph of Section 2.4, where it was noted that the posterior distribution $\{P(x|y)\}$ can be simulated using the Gibbs sampler, as suggested by Grenander (1983) and by Geman and Geman (1984). My dismissal of such an approach for routine applications was somewhat cavalier: purpose-built array processors could become relatively inexpensive if they are in sufficient demand. Furthermore, refined processing may be required only on a restricted region. Thus, let us suppose that, for 100 complete cycles say, images have been collected from the Gibbs sampler (or by Metropolis' method), following a "settling-in" period of perhaps another 100 cycles, which should cater for fairly intricate priors, bearing in mind the effect of informative records: I shall ignore storage problems. These 100 images should often be adequate for estimating properties of the posterior, such as (b) or some other local functional, and for making approximate associated confidence statements, as mentioned by Mr Haslett. Of course, this formally assumes the correctness of the posterior but some preliminary simulations, both by Dr Green and myself, suggest that the large-scale characteristics of the prior are relatively unimportant; otherwise, a globally modified prior should be used.

Dr Berman suggests producing a single image $\tilde{x}$ from noise-corrupted data, such that its spatial covariance properties are those of the prior. This seems wasteful of the records, though it is true that a criterion such as (b) may generate a quite atypical, usually rather smooth, scene, which is appropriate for estimating areas or volumes, say, but is visually unsatisfactory. An alternative is to use the Gibbs sampler to examine single images from the posterior distribution and hence obtain a better idea of a typical scene, given y. Naturally, one hopes that, in practice, the variability in the posterior is relatively small.

Mr Haslett's approach to (b) is of some interest, in a restrictive setting. However, the first two paragraphs of his discussion are severely flawed. In addition, his comparisons are of chalk with cheese: in particular, he allows himself large neighbourhoods, up to 11×11, with parameters estimated from the true image, whereas he condemns other methods to the most naive assumptions. Moreover, although he claims "a fair degree of accuracy in predicting the overall error rates", in fact these rates are often twice those he himself predicts.

Assessment

A number of discussants question the criteria by which one should judge success in restoration, given the true scene. In the paper, I chose the simplest measure, namely overall error rate, but clearly this can become a nonsense, for example with small regions against a uniform background. It would have been better at least to quote error rates for each colour separately, as does Dr Kershaw, and indeed this may be the most appropriate summary in the case of a crop inventory, say. In some contexts, the criteria of success might be very specific, as suggested by Professor Titterington; for example, in the identification of particular objects in a scene. However, as mentioned in Section 6, it would be very useful to have measures available which focus on colour boundaries: stochastic geometers and others please help! The preceding remarks apply to unordered colours but corresponding problems arise for grey-level scenes, with L1 and L2 providing obvious but sometimes suspect overall measures of discrepancy. When producing pictures for direct visual interpretation, rather than for the estimation of areas or volumes, say, the actual user must of course remain the final arbiter, a sentiment stressed by Verdi, Shepp and Kaufman (1985) in their work on tomography.

The pixels

Mixed pixels are discussed by Dr Jennison, Professor Mardia and Professor Switzer, and present a variety of problems. At one extreme, the true scene comprises single-colour patches that are generally large in comparison with the size of pixels: difficulties then occur only because colour boundaries intersect some of them. Naive processing may entirely misclassify such pixels because of the effect on their records, a defect with spatial processing should help to avoid. However, Dr Jennison's suggestion of splitting each pixel into quarters and using block *ICM* (or some other spatial method) to produce a refined estimate is extremely attractive, at least in principle. Unfortunately, it may present computational problems unless there are very few colours. A possible remedy is to assume that, at most, a single boundary intersects

any pixel, so that searches only involve pairs of colours: this device is used, somewhat differently, by Switzer (1983) and by Owen (1984).

The opposite extreme occurs when pixels are generally larger, perhaps much larger, than at least some of the objects of interest. Thus, Drs Campbell and Kiiveri have been concerned with classifying vegetation in the Australian bush, from Landsat data. In such cases, the scene itself can no longer be restored but it may yet be possible to estimate the proportion of each pixel covered by each of the colours, provided there are enough informative data channels. In a spatial Bayesian framework, the prior should address these continuous proportions: see Kent and Mardia (1986). However, for the geological data mentioned by Professor Mardia, most pixels are pure and the problem would seem to be of the first rather than the second type, in which case Dr Jennison's discrete approach is perhaps preferable.

The antithesis of pixel splitting is to amalgamate records and restore the image on a coarser, rather than a finer, grid. This may sometimes be sufficient in itself but, in methods demanding heavy computation, can also be used as a prelude to restoration at the pixel level (Gidas, 1985b). Note that, in computerised tomography, reconstruction is on an arbitrary grid and therefore, both size and shape are in the hands of the user: one possibility is to reconstruct the image on a honeycomb of hexagonal cells. Incidentally, tomographic reconstruction is usually carried out on a finer scale than the data really allow: for example, the maximum likelihood solution, implemented by Shepp and Vardi (1982), depends on the starting point of the associated *EM* algorithm.

Parameter values

There are many interesting contributions concerning the choice or estimation of parameters, both for the prior and for the likelihood. I have already commented on the value $\beta = 1.5$ in equation (6) and would like here to concentrate mainly on the estimation of parameters, particularly in the prior. There are at least two different situations. The first concerns the task of segmenting an image into regions of different known texture, such as wood, cork and plastic. In this case, abundant data are available from known scenes, so that multi-parameter models, over quite large neighbourhoods can be fitted to each texture using maximum pseudo-likelihood estimation. However, note that, in the real task, there may be awkward problems involving orientation of the textures. Proofs of asymptotic consistency and Normality of the pseudo-likelihood estimator, as lattice size increases, are given by Guyon (1986) and by Geman and Graffigne (1986). The latter proof is of particular interest because it does not depend on mixing conditions and is still valid in cases where maximum likelihood breaks down!

The second situation arises when the parameters and the scene have to be estimated concurrently, as in Section 5.1.2 of the paper. The number of such parameters then needs to be very small: indeed, Geman and Geman (1986) argue that a reparameterization of the prior can often be effected, so that it depends only on a single unknown value: I take this to be a point also made by Dr Kent (though I do not agree that my paper "tried to summarise spatial dependence in terms of a single parameter"!) and by Dr Possolo. In such cases, it is possible to implement maximum likelihood estimation, as in the calibration curve of Geman and Geman (1986) or by the neat one-step method suggested by Professor Künsch, which is available more generally. Thus, there is an alternative to the pseudo-likelihood scheme set out in Section 5.1.2 or in Dr Kiiveri's contribution. I have a few comments on the choice.

First, in the region of the parameter space for which maximum likelihood is consistent, it should also be more efficient, though the computational burden can be very heavy. Second, however, I should like to develop a point I made in Section 3, namely that different priors may be appropriate to different methods of restoration, though I shall confine myself here to different parameter values for the same model. This is illustrated in Section 5.4 and in the contribution from Drs Greig, Porteous and Seheult. The explanation is that different methods of restoration concentrate on different scales in the image and may therefore highlight different inadequacies in the prior. An immediate corollary is that the method of parameter estimation should also address the method of restoration: maximum pseudo-likelihood may be appropriate for *ICM*, since each is based on the local characteristics of the scene (for an exception to this general rule, see Section 4.2.1), but is likely to be inappropriate, if not disastrous (cf. $\beta = \infty$ in Section 5.4), for m.a.p.-like procedures. This also means that Professor Künsch's approach may break down, in practice, because of inadequencies in the prior. A last comment, in an area that is still very open, is that pseudo-likelihood estimation may be more relevant to the use of different parameter values in different parts of the scene, as proposed by Professor Titterington and Dr Clifford.

Practical applications

Until very recently, and in common with the great majority of statisticians, I have not had easy access

to sophisticated image processing equipment. Though this may explain the absence of any real examples in the paper, it does nothing to remedy the defect! I was therefore very grateful that so many discussants mentioned potential applications, from microscopy to satellite data, and from nuclear medicine to expert systems, though I recognize that many of the contributors are sceptical about the relevance of the approach I have advocated. I was pleased that Dr Kelly suggested real-time processing of moving images: the obvious context is in medical ultrasound imaging, as mentioned by Dr Kay and by Dr Scott, though ultrasonics can have some awkward reconstruction features.

The most encouraging point is that statisticians are becoming increasingly involved in image processing and allied activities. On a much wider front, I am in total sympathy with the sentiments expressed in Professor Titterington's and Dr Clifford's first paragraphs. Certainly in the U.K., we need more specialists in the new areas of science and technology, following the earlier examples of agriculture and medicine; in turn this will generate new and exciting theoretical developments in our subject.

REFERENCES IN THE DISCUSSION

Baum, L. E., Petrie, T., Soules, G. and Weiss, N. (1970) A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Ann. Math. Statist.*, **41**, 164–171.

Blödner, R., Mühlig, P. and Nagel, W. (1984) The comparison by simulation of solutions of Wicksell's corpuscle problem. *J. Microscopy*, **135**, 61–74.

Boyce, J. F., Feng, J. and Haddow, E. R. (1985) Relaxation labelling and the entropy of neighbourhood information. Preprint, Dept of Physics, King's College, London.

Brown, P. J. and Sundberg, R. (1986) Calibration and prediction diagnostics. Submitted for publication.

——(1987) Confidence and conflict in multivariate calibration *J. R. Statist. Soc.* B, **49**, in press.

Brown, P. J. (1982) Multivariate calibration (with Discussion). *J. R. Statist. Soc.* B, **44**, 287–321.

Critchley, F. and Ford, I. (1984) On the covariance of two non-central F random variables and the variance of the estimated linear discriminant function. *Biometrika*, **71**, 637–638.

——(1985) Interval estimation in discrimination: the multivariate normal equal covariance case. *Biometrika*, **72**, 109–116.

Faugeras, O. D. and Berthod, M. (1981) Improving consistency and reducing ambiguity is stochastic labelling: an optimization approach. *I.E.E.E. Trans. Pattern Anal. Machine Intell.*, **PAMI-3**, 412–423.

Ford, L. R. and Fulkerson, D. R. (1956) Maximal flow through a network. *Can. J. Math.*, **8**, 399–404.

Geman, S. (1984) Stochastic relaxation methods for image restoration and expert systems. In *Automated Image Analysis: Theory and Experiments* (D. B. Cooper *et al.*, eds). New York: Academic Press.

Geman, D. and Geman, S. (1986) Bayesian image analysis. In NATO ASI Series, vol. F20, *Disordered Systems and Biological Organization* (E. Bienenstock *et al.*, eds). Berlin: Springer-Verlag.

Geman, S. and Graffigne, C. (1986) A consistency theorem for Markov random fields. Technical Report, Brown University.

Geman, S. and Hwang, C-R. (1986) Diffusions for global optimization. *SIAM J. Control and Optimization*, **24**, in press.

Gidas, B. (1985a) Global optimization via the Langevin equation. *I.E.E.E. Trans.*, pp. 774-778.

——(1985b) A renormalization group approach to image processing problems. Technical Report, Brown University.

Guyon, X. (1986) Estimation d'un champ par pseudo-vraisemblance conditionelle. *Actes 6eme Rencontre Franco-Belge de Statisticiens* (to appear).

Haslett, J. and Horgan, G. (1985) Spatial discriminant analysis—A linear discriminant function for the black/white case. Paper presented at the Edinburgh Meeting on Pattern Recognition, 1985, and submitted for publication.

Kalayeh, H. M. and Landgrebe, D. A. (1984) Adaptive relaxation labelling. *I.E.E.E. Trans. Pattern Anal. Machine Intell.*, **PAMI-6**, 369–372.

Kiiveri, H., Speed, T. P. and Carlin, J. B. (1984) Recursive causal models. *J. Austr. Math. Soc.* A, **36**, 30–52.

Mardia, K. V. and Hainsworth, T. (1986) A spatial thresholding method for image segmentation. To appear.

O'Hagan, A. (1978) Curve fitting and optimal design for regression (with Discussion). *J. R. Statist. Soc.* B, **40**, 1–42.

Owen, A. (1984) A neighbourhood-based classifier for LANDSAT data. *Can. J. Statist.*, **12**, 191–200.

Pearl, J. (1986) A constraint-propagation approach to probabilistic reasoning. In *Uncertainty and Artificial Intellience* (L. N. Kanal and J. Lemmer, eds). Amsterdam: North-Holland.

Phelps, M. E., Mazziota, J. C. and Schelbert, H. R. (1986) Positron emission tomography and autoradiography. Principles and applications for heart and brain. Raven Press.

Picard, J. C. and Ratliff, H. D. (1975) Minimum cuts and related problems. *Networks*, **5**, 357–370.

Poso, S., Hame, T. and Paananen, R. (1984) A method of estimating the stand characteristics of a forest compartment using satellite imagery. *Silva Fennica*, **18**, 261–292.

Possolo, A. (1986) Estimation of binary Markov random fields. Technical Repot No. 77, Dept of Statistics, University of Washington.

Richards, J. A., Landgrebe, D. A. and Swain, P. H. (1981) On the accuracy of pixel relaxation labelling. *I.E.E.E. Trans. Systems, Man and Cybernetics*, **SMC-11**, 303-309.

Russ, J. C. and Russ, J. C. (1984) Image processing in a general purpose microcomputer. *J. Microscopy*, **135**, 89–102.

Shepp, L. A. and Vardi, Y. (1982) Maximum likelihood reconstruction in positron emission tomography. *I.E.E.E. Trans. Medical Imaging*, **1**, 113–122.

Silverman, B. W. (1985) Some aspects of the spline smoothing approach to non-parametric regression curve fitting (with Discussion). *J. R. Statist. Soc.* B, **47**, 1–52.

Silvey, S. D., Titterington, D. M. and Torsney, B. (1978) An algorithm for optimal designs on a finite design space. *Commun. in Statist.*, A, **7**, 1379–1389.

Spiegelhalter, D. J. (1986a) A statistical view of uncertainty in expert systems. In Artificial Intelligence and Statistics (W. Gale, ed.). Reading, Mass.: Addison Wesley.

Titterington, D. M. (1985) Common structure of smoothing techniques in statistics. *Int. Statist. Rev.*, **53**, 141–170.

Torsney, B. (1977) In discussion of Dempster, Laird and Rubin's paper. *J. R. Statist. Soc.* B, **39**, 26–27.

——(1981) Ph. D. Thesis, University of Glasgow.

——(1983) A moment inequality and monotonicity of an algorithm. In *Lecture Notes in Economics and Mathematical Systems 215* (Proceedings of the International Symposium on Semi-infinite Programming and Applications at the University of Texas at Austin, 1981) (K.O. Kortanek and A. V. Fiacco, eds), pp. 249–260.

——(1986b) Probabilistic reasoning in predictive expert systems. In Uncertainty and Artificial Intelligence (L. N. Kanal and J. Lemmer, eds). Amsterdam: North-Holland.

Weiss, S. M. *et al.* (1978) A model-based method for computer-aided medical decision-making. *Artif. Intell.*, **11**, 145–172.

Wermuth, N. and Lauritzen, S. L. (1983) Graphical and recursive models for contingency tables. *Biometrika*, **70**, 537–552.

Vardi, Y., Shepp, L. A. and Kaufman, L. (1985) Emission tomography. *J. Amer. Statist. Ass.*, **80**, 8–37.

Verhagen, A. M. W. (1977) A three parameter isotropic distribution of atoms and the hard-core square lattice gas. *J. Chem. Phys.*, **67**, 5060–5065.

Young, T. Y. and Calvert, T. W. (1974) *Classification, Estimation and Pattern Recognition.* New York: Elsevier.

Zucker, S. W., Krishnamurthy, E. V. and Haar, R. L. (1978) Relaxation processes for scene labelling: convergence, speed and stability. *I.E.E.E. Trans. Systems, Man and Cybernetics*, **1**, 41–48.

Zucker, S. W. Leclerc, Y. G. and Mohammed, J. L. (1981) Continuous relaxation and local maxima selection: conditions for equivalance. *I.E.E.E. Trans. Pattern Anal. Machine Intell.*, **2**, 117–127.