Exact Maximum A Posteriori Estimation for Binary Images

D. M. Greig; B. T. Porteous; A. H. Seheult

Journal of the Royal Statistical Society. Series B (Methodological), Vol. 51, No. 2. (1989), pp. 271-279.

Stable URL:
http://links.jstor.org/sici?sici=0035-9246%281989%2951%3A2%3C271%3AEMAPEF%3E2.0.CO%3B2-5

Journal of the Royal Statistical Society. Series B (Methodological) is currently published by Royal Statistical Society.

J. R. Statist. Soc. B (1989)
51, *No.* 2, *pp.* 271–279

Exact Maximum *A Posteriori* Estimation for Binary Images

By D. M. GREIG, B. T. PORTEOUS and A. H. SEHEULT†

University of Durham, UK

[Received June 1987. Final revision September 1988]

SUMMARY

In this paper, for a degraded two-colour or binary scene, we show how the image with maximum *a posteriori* (MAP) probability, the MAP estimate, can be evaluated exactly using efficient variants of the Ford–Fulkerson algorithm for finding the maximum flow in a certain capacitated network. Availability of exact estimates allows an assessment of the performance of simulated annealing and of MAP estimation itself in this restricted setting. Unfortunately, the simple network flow algorithm does not extend in any obvious way to multicolour scenes. However, the results of experiments on two-colour images suggest that, in general, simulated annealing, according to *practicable* 'temperature' schedules, can produce poor approximations to the MAP estimate to which it converges.

Keywords: BAYESIAN METHODS; CAPACITATED NETWORKS; FORD–FULKERSON ALGORITHM; IMAGE PROCESSING; ITERATED CONDITIONAL MODES; MARKOV RANDOM FIELDS; MAXIMUM *A POSTERIORI* ESTIMATION; SIMULATED ANNEALING

1. INTRODUCTION

There has been considerable interest in Bayesian methods for image analysis: see, for example, Geman and Geman (1984) and Besag (1986). In this paper, we show for a degraded binary image how the maximum *a posteriori* (MAP) estimate of the true scene can be found *exactly*. This allows limited assessment of MAP estimation and simulated annealing approximations to it. We also include comparisons with the method of iterated conditional modes (ICM); see Besag (1986).

With x_i denoting the category or value of pixel i in the image $x = (x_1, \ldots, x_n)$, a Bayesian formulation specifies an *a priori* distribution $p(x)$ over all allowable images. Usually, $p(x)$ is taken to be a locally dependent Markov random field (MRF), a convenient model for quantifying the belief that the unknown true scene x^* consists of, for example, large homogeneous patches, or smoothly varying grey levels which occasionally change level discontinuously. With $y = (y_1, \ldots, y_n)$ denoting the observed records of x^*, the likelihood $l(y|x)$ of any image x is combined with $p(x)$, in accordance with Bayes's theorem, to form an *a posteriori* distribution $p(x|y) \propto l(y|x)p(x)$. The MAP estimate of x^* is that image $\hat{x}$ which maximizes $p(x|y)$.

However, direct calculation of $\hat{x}$ is in general computationally prohibitive and therefore, as an approximation, Geman and Geman (1984) have proposed the use of simulated annealing, a stochastic relaxation algorithm which, for an appropriate 'temperature' schedule, converges to $\hat{x}$. In the special case of binary images, Greig *et al.* (1986) indicated how $\hat{x}$ can be found exactly. Very limited experimentation then suggested that simulated annealing can produce poor approximations, with $\hat{x}$ tending to be smoother than the annealing approximation to it.

† *Address for correspondence*: Department of Mathematical Sciences, University of Durham, Science Laboratories, South Road, Durham, DH1 3LE, UK.

 0035-9246/89/51271 $2.00

In Section 2, MAP estimation for binary images is reformulated as a minimum cut problem in a certain capacitated network, and the classic Ford–Fulkerson algorithm can then be used to find $\hat{x}$ exactly. In Section 3, the results of extensive experimentation are reported. A much faster version of the Ford–Fulkerson algorithm, which exploits the local connectedness of the associated image network, is described in Section 4.

2. EXACT MAXIMUM *A POSTERIORI* ESTIMATION FOR BINARY IMAGES

In this Section, we present details of a network reformulation of the MAP estimation problem for binary images. Each pixel is to be classified into one of two unordered colours or categories, which we call white and black and code as $x_i = 0$ and $x_i = 1$ respectively. Following Besag (1986), we adopt the following assumptions.

The records $y_1, \ldots, y_n$ are conditionally independent given x, and each has known conditional density function $f(y_i|x_i)$, dependent on x only through x_i. Thus, the likelihood function for x may be written

$$l(y|x) = \prod_{i=1}^{n} f(y_i|x_i) = \prod_{i=1}^{n} f(y_i|1)^{x_i} f(y_i|0)^{1-x_i}.$$

The prior distribution $p(x)$ is modelled as a pairwise interaction MRF of the form

$$p(x) \propto \exp\left[\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\beta_{ij}\{x_i x_j + (1 - x_i)(1 - x_j)\}\right]$$

where $\beta_{ii} = 0$ and $\beta_{ij} = \beta_{ji} \geqslant 0$; in the strict inequality, we call i and j neighbours. Such a distribution is a special case of the general pairwise interaction MRF in equation (9) of Besag (1986); in fact, the results of this paper hold for any binary pairwise MRF provided that, in Besag's (1986) notation, $G_{ij}(x_i, x_j)$ is non-negative for $x_i = x_j$ and non-positive otherwise. If $\beta_{ij} = \beta$, if i and j are neighbours, and $\beta_{ij} = 0$ otherwise, then $p(x) \propto \exp(\beta v)$, where v is the number of neighbour pairs with like colours.

Thus, apart from an additive constant, $\ln p(x|y)$ can be written as

$$L(x|y) = \sum_{i=1}^{n}\lambda_i x_i + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\beta_{ij}\{x_i x_j + (1 - x_i)(1 - x_j)\}$$

where $\lambda_i = \ln\{f(y_i|1)/f(y_i|0)\}$, a log-likelihood ratio at pixel i. The MAP estimate is that image $\hat{x}$ which maximizes L. There are 2^n possible values of L and n may be of the order 256×256, making direct search for $\hat{x}$ infeasible. Fortunately, there are efficient algorithms for maximizing L, which we now describe.

Consider a capacitated network comprising $n + 2$ vertices, being a source s, a sink t and the n pixels. There is a directed edge (s, i) from s to pixel i with capacity $c_{si} = \lambda_i$, if $\lambda_i > 0$; otherwise, there is a directed edge (i, t) from i to t with capacity $c_{it} = -\lambda_i$. There is an undirected edge (i, j) between two internal vertices (pixels) i and j with capacity $c_{ij} = \beta_{ij}$ if the corresponding pixels are neighbours.

For any binary image $x = (x_1, \ldots, x_n)$ let $B = \{s\} \cup \{i: x_i = 1\}$ and $W = \{i: x_i = 0\} \cup \{t\}$ define a two-set partition of the network vertices and put

$$C(x) = \sum_{k \in B}\sum_{l \in W} c_{kl}.$$

The set of edges with a vertex in B and a vertex in W is called a *cut* and $C(x)$ is called the *capacity* of the cut.

It is readily seen that $C(x)$ may be written

$$C(x) = \sum_{i=1}^{n} x_i \max(0, -\lambda_i) + \sum_{i=1}^{n} (1 - x_i)\max(0, \lambda_i) + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \beta_{ij}(x_i - x_j)^2$$

which differs from $-L(x|\, y)$ by a term which does not depend on x; see also Picard and Ratliff (1975). Moreover, as Ford and Fulkerson (1962) show, the minimum of $C(x)$ is the maximum flow through the network from source to sink subject to the edge capacities, and they give an efficient algorithm for solving this problem. A corresponding cut is called a *minimum cut*. Thus, maximizing $L(x|\, y)$ is equivalent to finding the minimum cut in the network: in the MAP estimate pixels are black if they are on the source side of the minimum cut and white otherwise.

3. SOME NUMERICAL COMPARISONS BETWEEN EXACT MAXIMUM *A POSTERIORI* ESTIMATION, SIMULATED ANNEALING AND ITERATED CONDITIONAL MODES

Two synthetic binary scenes were used in our experiments and are displayed in Fig. 4a of Besag (1986) and Fig. 1(a). Different methods were applied to records created by corrupting the synthetic scenes with random noise from known distributions and experiments were replicated using independent realizations from these same noise distributions. The simple prior distribution $p(x) \propto \exp(\beta v)$ was used throughout, where v denotes the number of neighbour pairs with like colours for the neighbourhood system comprising the eight adjacencies of each pixel, save for the obvious boundary modifications. The effect of varying β was investigated and different annealing schedules were compared. In this study β was held constant throughout the eight iterations used in each application of ICM.

Records for the first image were created by adding independent Gaussian noise with variance 0.9105 to the 0–1 version of the 88 × 100 white-and-black scene displayed in Fig. 4a of Besag (1986). This leads to a 30% expected misclassification rate for the maximum likelihood classifier. The estimators studied were applied to each of five replicates of records for five values of β. The results are summarized in Table 1 and will be discussed in detail here.

Records for the second image were created by applying a binary channel with 25% error rate to the 64 × 64 white-and-black scene in Fig. 1(a), leading to a 25% expected misclassification rate for the maximum likelihood classifier; an example is displayed in Fig. 1(b). The estimators studied were applied to each of five replicates of records for three values of β. The results are summarized in Table 2, and some of the corresponding image estimates are displayed in Fig. 1.

For each β and for each estimator considered, the replicate mean and standard deviation of the misclassification rate are given in the upper sections of Tables 1 and 2. For any replicate, consider the deviation of log-posterior probability of an estimate from the log-posterior probability of the corresponding MAP estimate. Replicate means and standard deviations for these log-posterior deviations are given in the lower sections of Tables 1 and 2, except for MAP estimation where because of the omitted additive constant only the replicate standard deviations are comparable.

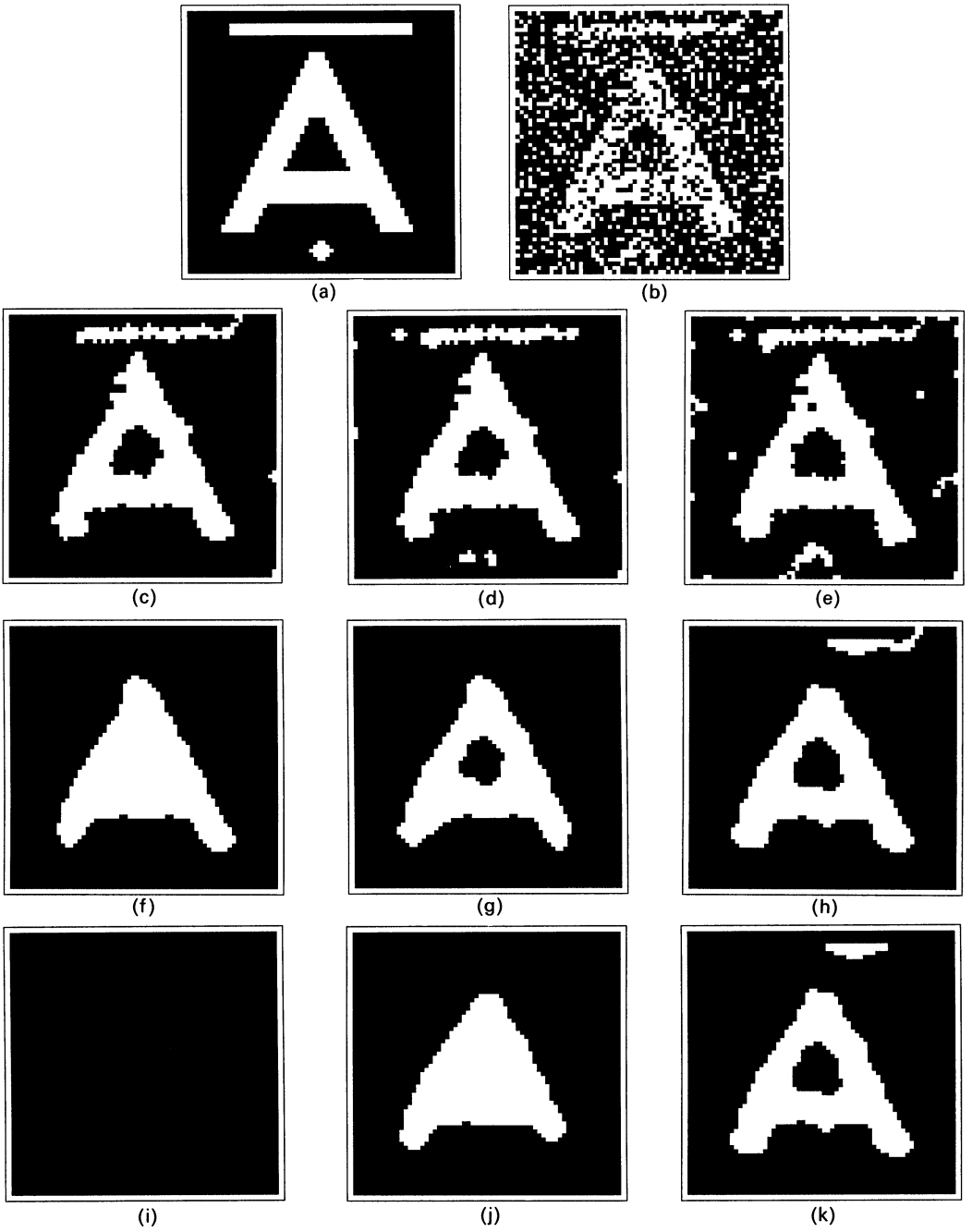

Fig. 1. (a) True 64 × 64 binary scene; (b) true scene corrupted by a binary channel with 25% error rate; (c) exact MAP estimate ($\beta = 0.3$); (d) simulated annealing estimate with geometric schedule $A\rho^{k-1}$ ($k = 1, \ldots, K$), with $A = 2/\ln 2$, $\rho = 0.99$ and $K = 565$ ($\beta = 0.3$); (e) ICM estimate ($\beta = 0.3$); (f) exact MAP estimate ($\beta = 0.7$); (g) simulated annealing estimate with geometric schedule $A\rho^{k-1}$ ($k = 1, \ldots, K$), with $A = 2/\ln 2$, $\rho = 0.99$ and $K = 565$ ($\beta = 0.7$); (h) ICM estimate ($\beta = 0.7$); (i) exact MAP estimate ($\beta = 1.1$); (j) simulated annealing estimate with geometric schedule $A\rho^{k-1}$ ($k = 1, \ldots, K$), with $A = 2/\ln 2$, $\rho = 0.99$ and $K = 565$ ($\beta = 1.1$); (k) ICM estimate ($\beta = 1.1$)

TABLE 1

Comparison of MAP estimation, simulated annealing and ICM for various values of β based on five replications of Gaussian noise superimposed on the 88 × 100 true scene depicted in Fig. 4a of Besag (1986)†

β	MAP	Simulated annealing						ICM
		Logarithmic				Geometric		
		C 2	2	4	ρ 0·95	0.99	0.995	
		K 5000	750	5000	K 112	565	1131	
Misclassification rate								
0.0	30.2	—	—	—	—	—	—	30.2
	0.9	—	—	—	—	—	—	0.9
0.3	5.5	6.1	6.3	8.1	5.7	5.6	5.4	7.6
	0.5	0.3	0.4	0.5	0.5	0.5	0.6	0.7
0.5	6.7	6.0	5.8	7.0	5.6	5.6	5.8	6.4
	0.8	0.8	0.9	0.6	0.3	0.8	0.8	0.3
0.7	9.5	7.7	7.3	8.5	6.6	7.1	7.7	7.0
	1.3	0.2	0.4	0.6	0.4	0.8	0.5	0.4
0.9	16.8	9.7	9.5	11.4	8.0	9.7	9.2	7.7
	3.3	1.4	1.3	0.7	0.7	0.6	1.0	0.5
1.1	27.1	12.2	11.7	14.2	9.5	10.8	12.1	8.3
	2.6	1.1	0.9	1.4	0.4	0.2	1.0	0.4
Log-posterior probability deviation								
0.0	—	—	—	—	—	—	—	—
	40‡	—	—	—	—	—	—	40‡
0.3	—	47	86	290	4	2	1	75
	66‡	5	10	22	2	2	0	14
0.5	—	21	33	100	25	10	5	153
	73‡	3	5	16	15	5	3	11
0.7	—	39	55	68	69	29	21	271
	79‡	20	21	20	21	11	12	22
0.9	—	91	130	83	158	79	95	452
	59‡	40	41	31	34	23	46	81
1.1	—	190	221	171	335	224	200	736
	53‡	41	52	53	58	45	48	96

†Each entry pair comprises a replicate mean (upper number) and a replicate standard deviation (lower number) of percentage misclassification rate or of log-posterior probability deviation from corresponding MAP estimate values.
‡Standard deviation of log-posterior probabilities for MAP estimates.

For simulated annealing, logarithmic schedules of the form $C/\ln(1 + k)$ and geometric schedules of the form $A\rho^{k-1}$ $(k = 1, \ldots, K)$ were used for both images. The value $A = 2/\ln 2$ was used in every geometric schedule for both experiments, giving the same starting temperature as a logarithmic schedule with $C = 2$. For each geometric schedule, K was chosen so that the final temperature was just less than 0.01.

Geometric schedules are often used in applications of simulated annealing to solve general complex optimization problems; see, for example, Kirkpatrick *et al.* (1983). An obvious advantage of geometric schedules over logarithmic schedules is that temperatures close to zero can be reached from a high starting temperature in considerably fewer iterations. However, in theory, a geometric schedule is too fast and there is the risk that a local maximum may be reached from which the annealing algorithm cannot escape.

TABLE 2

Comparison of MAP estimation, simulated annealing and ICM for various values of β based on five replications of binary noise applied to the 64 × 64 true scene depicted in Fig. 1(a)†

β	*MAP*	*Simulated annealing*				*ICM*
		Logarithmic		*Geometric*		
		C 2	ρ *0.95*	*0.99*	*0.995*	
		K 750	*K 112*	*565*	*1131*	
Misclassification rate						
0.0	24.8	—	—	—	—	24.8
	0.5	—	—	—	—	0.5
0.3	5.2	6.6	5.3	5.3	5.4	6.9
	0.5	0.7	1.0	0.6	0.8	1.0
0.7	9.6	7.9	7.2	7.2	7.2	6.4
	1.6	0.4	0.5	0.3	0.7	0.8
1.1	22.8	10.4	9.1	10.6	11.1	6.3
	0.0	1.4	0.8	1.5	1.4	0.8
Log-posterior probability deviation						
0.0	—	—	—	—	—	—
	0‡	—	—	—	—	0‡
0.3	—	39	3	1	1	19
	23‡	6	2	1	0.45	6
0.7	—	20	31	17	16	94
	39‡	12	21	8	9	35
1.1	—	59	123	63	55	281
	46‡	30	50	32	29	98

†Each entry pair comprises a replicate mean (upper number) and a replicate standard deviation (lower number) of percentage misclassification rate or of log-posterior probability deviation from corresponding MAP estimate values.
‡Standard deviation of log-posterior probabilities for MAP estimates.

We now discuss the results in Tables 1 and 2; as the general pattern is similar for both images, we focus attention mainly on Table 1, referring to Table 2 only when differences arise.

Mean misclassification rate is generally a U-shaped function of β for all estimators: for small values of β estimators tend to undersmooth, whereas for large values of β they tend to oversmooth. Rates increase with β most rapidly for MAP estimation and least rapidly for ICM, whereas the increase for simulated annealing is intermediate. Thus, MAP estimation is very sensitive to changes in the specification of the prior, whereas ICM is generally robust to such changes. Rates for simulated annealing underestimate those for MAP, particularly for higher values of β. As we would expect, rates for geometric schedules tend to increase with ρ towards the rate for the corresponding MAP estimate.

The standard deviation of the misclassification rate shows little variation for simulated annealing and ICM, though it tends to be greatest for large values of β. However, the standard deviation for MAP estimation increases, then decreases, with increasing β because, for large β, the MAP estimate tends to be one colour.

From Table 1, we see that mean log-posterior probability deviation increases with β for both ICM and for the three geometric annealing estimates but is U shaped for the three logarithmic annealing estimates. The geometric schedules give results that are very close to the MAP estimates for β between 0.3 and 0.5 and are better than the

corresponding logarithmic schedules, presumably because in this range there is less chance of becoming trapped in a local maximum of $p(x|y)$ than for larger values of β. However, for large values of β the opposite appears to be the case, suggesting that in this range the geometric decrease is too fast. These observations are supported by the corresponding increase in standard deviation of log-posterior deviation and are related to results of experiments reported to us by Ripley (1988) which indicate that, for the larger values of β considered, geometric schedules produce essentially different restorations according to which seed is used to initiate the annealing algorithm. Thus, the opportunity for the simulated annealing algorithm to become trapped in a local maximum of $p(x|y)$ increases with increasing β especially if the temperature is allowed to decrease too rapidly.

The importance of the rate of decrease in temperature is illustrated by the geometric schedules where mean log-posterior probability deviation decreases with increasing ρ. The results for the logarithmic schedules indicate the influence of starting temperature. Thus, for large β many iterations of a slow schedule starting at a high temperature will be required to produce a good simulated annealing approximation to an MAP estimate, whereas for small β fewer iterations of a faster schedule starting at lower temperature should be sufficient.

We close this section with a few comments on MAP estimation itself. Our results indicate that exact MAP estimation can be extremely sensitive to specification of the prior distribution. Thus, in image restoration and classification, for example, we have seen that such misspecification can lead to drastic oversmoothing and misclassification. However, there may be other contexts, such as boundary detection, where MAP estimation is the preferred method; see, for example, Geman *et al.* (1988).

4. MORE EFFICIENT VARIANT OF THE FORD–FULKERSON ALGORITHM

The first set of experiments described in Section 3 is based on an image of size 88×100 and, for definiteness, we now consider $\beta = \frac{1}{3}$. On an Amdahl 470 computer, the basic algorithm required 3000 s of central processor unit (CPU) time to find the MAP estimate. However, a variant of the basic algorithm has reduced this to 250 s.

The variant is as follows: partition the image into connected subimages and then calculate the MAP estimate for each subimage separately. This can be interpreted as finding the maximum flow through the network, but under the imposed constraints that no flow is allowed across subimage boundaries.

At the next stage, we relax some of the constraints that no flow is allowed across subimage boundaries. Corresponding subimages are thus amalgamated to form a new set of larger subimages and the MAP estimate for each of these is then obtained. This procedure continues until, at the last stage, we are left with the MAP estimate of the complete image.

To achieve the twelvefold reduction in CPU time mentioned, the 88×100 image was first partitioned into a 16×16 array of roughly equal-sized rectangular subimages. Once the individual MAP estimates had been obtained, neighbouring subimages were amalgamated in fours to leave a new 8×8 array of subimages. This procedure was continued until the MAP estimate for the complete image was obtained.

This choice of partitioning is unlikely to be optimal, but compared with the unpartitioned version any sensible choice of partition will lead to a substantial reduction in CPU time.

5. DISCUSSION

This paper has described a method for exact MAP estimation of binary images. In this simple case the results can be compared directly with those using simulated annealing, and the efficacy of that method in producing MAP estimates can be assessed. Our results indicate that simulated annealing, applied with practicable schedules, does not necessarily produce a good approximation to an MAP estimate. However, our experimental results suggest that good approximations are more likely for smaller values of the parameter β, and in such cases geometric schedules outperform logarithmic schedules. An informal explanation of this behaviour is that the amount of smoothing required to obtain an MAP estimate increases with β and, as Besag (1986), p. 298, has conjectured, the simulated annealing algorithm then appears to become 'increasingly bogged down by local maxima' resulting in undersmooth approximations.

Since an exact MAP estimate is now available, some insights into its value as a method of image restoration can be obtained. Our results emphasize the crucial importance of the assumed prior distribution; in particular, we demonstrate that, as β increases, the global properties of the prior distribution very rapidly dominate the likelihood contribution to the posterior distribution. We conjecture that corresponding multicolour estimates will behave similarly.

Assuming the model of Section 2, obtaining a binary MAP estimate has been shown to be equivalent to optimizing a particular quadratic function of 0–1 variables; it is this problem which is amenable to a network flow solution. Any attempt to incorporate line sites, as in Geman and Geman (1984), or to preserve certain global aspects of the true scene, considered by Green (1986), will in general render the network method inapplicable. Similarly, the multicolour problem cannot be directly dealt with by this method: though it can be treated as a generalized minimum cut problem, there seems to be no corresponding network formulation.

ACKNOWLEDGEMENTS

We are extremely grateful to Julian Besag for suggesting the problem and for his insightful comments on an earlier version of the paper. We also thank Peter Green for his encouragement and helpful comments and Brian Ripley for his active interest in our work. Most of our experimentation was performed on a Sun 3/160 workstation, the funds for which were provided by a Science and Engineering Research Council grant.

REFERENCES

Besag, J. E. (1986) On the statistical analysis of dirty pictures (with discussion). *J. R. Statist. Soc.* B, **48**, 259–302.

Ford, L. R. and Fulkerson, D. R. (1962) *Flows in Networks*. Princeton: Princeton University Press.

Geman, D. and Geman, S. (1984) Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans. Pattn Anal. Mach. Intell.*, **6**, 721–741.

Geman, D., Geman, S., Graffigne, C. and Dong, P. (1988) Boundary detection by constrained optimization. Unpublished.

Green, P. J. (1986) Discussion on On the statistical analysis of dirty pictures (by J. E. Besag). *J. R. Statist. Soc.* B, **48**, 284–285.

Greig, D. M., Porteous, B. T. and Seheult, A. H. (1986) Discussion on On the statistical analysis of dirty pictures (by J. E. Besag). *J. R. Statist. Soc.* B, **48**, 282–284.

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

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

Ripley, B. D. (1988) Some experiments with simulated annealing for MAP restoration of images. *Personal communication.*