REFERENCES

Diggle, P. J., Fiksel, T., Grabarnik, P., Ogata, Y., Stoyan, D. and Tanemura, M. (1990). On parameter estimation for pairwise interaction point processes, Tech. Report (submitted).

Särkkä, A. (1990). Applications of Gibbs point processes: pseudo-likelihood estimation method with comparisons, Reports from the Department of Statistics, University of Jyväskylä 10/1990.

REJOINDER

JULIAN BESAG*

Department of Statistics GN-22, University of Washington, Seattle, WA 98195, U.S.A.

I begin by thanking all the discussants for their very valuable comments. Many issues have been raised and I cannot hope to answer them all. I particularly thank those discussants who have provided new analyses of one or both types of example presented in the paper. I must also explain immediately that this reply represents my own views and not necessarily those of my co-authors, both of whom made important contributions to the paper, as part of their graduate studies. I hope this does not appear discourteous but there are a number of logistical constraints, partly brought about by the fact that the three of us are in separate countries and have not met nor worked together for some considerable time now, and partly because of an imminent deadline.

Background

In order to set the paper in context, it may be helpful to acquaint general readers with its background. In the first instance, a version was written for the "Symposium on the Analysis of Statistical Information", held in Tokyo in December 1989, and appears in the proceedings of that meeting. Subsequently, Professor Kitagawa very kindly invited us to submit a modified account, as a discussion paper, to *Ann. Inst. Statist. Math.* The main modification was to be the inclusion of at least one example relating to the mapping of disease. The version that appears in the conference proceedings omits examples from Section 4, though the spoken presentation did include all three. The reason for the omission was that it was not yet clear that Bayesian mapping was at a stage to be put forward as a tool for

* Now at Department of Mathematics and Statistics, University of Newcastle upon Tyne, Newcastle upon Tyne, NE1 7RU, U.K.

routine use in epidemiology. That remains true today, in my view at least, and I have resisted, on several occasions, requests to produce 'Moff-the-shelf" analyses of cancer registry data. Much still remains to be done. As regards the archaeological example, this was carried out one afternoon in my office in Durham, mainly for fun, using an image analysis program modified to take account of missing values. At the time, it was not intended for publication, nor have I communicated the results to the archaeologists who collected the data. Clearly then, the examples have achieved a rather more exhalted status than was originally intended!

This said, I hope the analyses provide useful illustrations of a line of research in spatial statistics that may prove profitable in the future. The constructive comments of the discussants are extremely helpful in this respect. At the very least, the methods will require considerable refinement before they can be considered as everyday components of the spatial statistics toolkit. In fact, it is interesting to note, as Adrian Raftery and Jeffrey Banfield point out in regard to the Gibbs sampler, there are quite conventional Bayesian problems, especially those concerned with hierarchical formulations, that have benefited from an image analysis interpretation. In addition, the methods have also found useful application in some other areas, including pedigree analysis (Sheehan (1990), Thomas (1991)), speech recognition (Lippman (1991)), neurophysiology (Fredkin and Rice (1991)) and classical statistical inference (Geyer and Thompson (1991), Besag and Clifford (1989, 1991)).

Despite the reservations expressed above, it is not my intention to excuse myself from responsibility concerning the particular examples in the paper and I shall try to answer the criticisms as best I can. I stand by the results of the analyses, though obviously there are improvements that could be made to the methodology and its implementation. A good example of this occurs in Section 3 of the paper, which has at least one unsatisfactory methodological aspect, also apparent in Besag ((1986), Section 5.1.2), and which I shall modify during the course of my reply; however, by good fortune, the correction has little effect when applied to the archaeological data and this is also the case in Besag ((1986), Section 5.1.3).

I have grouped my responses under headings. Some of my remarks are extremely vague and I hope that readers with more complete understanding will be able either to substantiate or refute them with better authority. I would be glad to hear from anyone who has results or suggestions for further work.

The role of the prior distribution

A theme that runs through much of the discussion concerns the role of the prior distribution in the Bayesian analysis of spatial data. The issues at stake would seem to include the following: (i) the purpose of the prior in relation to the observed records; (ii) the effect of different hyperparameter values; (iii) the properties of the resulting posterior distribution; (iv) the criterion on which a point estimate of an "image" should be based; (v) the estimation of hyperparameters; (vi) the relevance and robustness of inferences made from the posterior distribution. I shall say a little about each of these issues, with some emphasis on the

particular examples in Sections 3 and 4; any attempt at sweeping generalizations would be inappropriate.

As regards the purpose of the prior, it is stated in Section 2 of the paper, "We do not necessarily require that typical realizations of $\{p(x)\}$ should resemble the true scene but that the distribution should at least support the local regularities that are believed to exist." I think we should have added "and that are partially evident in the observed records." For example, I would have had serious misgivings in Section 3 had there been more than a few missing data points. The aim is to massage the data rather than to trample on or invent them. Indeed, I have suggested elsewhere (Besag (1986)) that the c-colour Potts model,which becomes the Ising model used in Section 3 when $c = 2$, can be thought of as a representation of "prior ignorance" about a patchy scene. This needs some qualification but first there is the simple problem of semantics. I have long objected to extravagant use of the term "model" in spatial statistics when no attempt at proper modelling is made. Thus, in Besag (1974, 1975), I used the terms "scheme" and 'Mprescription" but to little avail and so I have reluctantly reverted to standard terminology. There are other situations where detailed models are highly appropriate, as for example in the use of deformable templates in structural image restoration (Chow *et al.* (1988), Amit *et al.* (1991)) or in texture analysis (Geman and Graffigne (1987)), though, even here, simple, flexible priors may be preferable (Geman *et al.* (1990)). Whichever the case, goodness-of-fit tests (cf. the remark by Dietrich Stoyan), or something less formal, are likely to be of importance only in choosing between different models (cf. Yosihiko Ogata's comments) or as a means of discrimination between textures, as in Geman *et al.* (1990). As in almost all Bayesian analysis, it is the data that are of paramount importance. Nonetheless, an exact goodness-of-fit test for the Ising model is given in Besag and Clifford (1989).

However, the issue raised by several discussants concerning Ising and Potts priors is not one of semantics but refers to the capacity of such models to produce "long-range order"; that is, for moderate and large values of β, pixels arbitrarily far apart have values that are positively correlated and, on an infinite array, this implies the existence of infinite, connected (in terms of neighbourhood), single-colour patches. On the other hand, small values of β do not have sufficient local dependence to produce a blobby structure. Although the obvious conclusion is that one needs to look elsewhere for useful priors, I shall attempt to argue the contrary, though without total conviction! I should add that my argument is not an "after the event" one, an attempt to extricate myself from an embarrassing situation, since I have warned about long-range order and the slowness of Metropolis or Gibbs sampler convergence for a good many years now. Whether I should have heeded my own warnings more carefully remains to be seen!

My first premise is that the important properties of the prior, the ones on which we need to concentrate, are those that are to any appreciable extent inherited by the posterior. This is presumably not in question but it does pose a very difficult problem. In physical terms, the observed records have the effect of applying an external magnetic field to the system. Infinite single-colour patches will still occur on the infinite array under certain conditions and need not all be of the same colour, since the external field is not uniform. Thus, in principle, the posterior

distribution may be concentrated on restorations that are almost one-colour or on segmentations that are in strong conflict with the observations and have no semblance of reality. Moreover, the maximum probability (MAP) estimate may produce such a result, even when the posterior distribution is more diffuse! Indeed, this occurs with the Potts prior for *small* positive values of β, provided the records are sufficiently uninformative; clearly then this is not the product of long-range order. It is this additional setback that I take to be the "real message" in the paper by Greig *et al.* (1989) and it is the basis of my frequent arguments (e.g. Besag (1986, 1989)) against the use of MAP, unless the posterior distribution is known to be unimodal (as in the posterior density of u and v, given y, κ and λ, in Section 4 of the paper). In fact, it is the general multi-modality that drives my second, more risky premise, that primarily I am interested in the posterior distribution only for values of x for which the likelihood is relatively large. Thus, I do not really want to sample the complete posterior distribution in the presence of multi-modality. The premise is a not too distant cousin of Savage's Principle of Precise Measurement, as I shall outline below.

As has been made clear, the prior used in Section 3 is a fairly crude representation of my initial beliefs about the *local* regularities in the true scene. A more complete description would have included the fact that *a priori* I believed the data to be meaningful, in that they would provide a fair reflection of the true state of nature. How could I have catered for this? Well, in the spirit of empirical Bayes, I might, for example, have modified the prior so as to penalize images whose colour frequencies differed markedly from those of the maximum-likelihood classifier, in the manner of Green (1986). The effect of this would be to leave the posterior distribution virtually unaltered, apart from rescaling, conditional on being in the required neighbourhood. *Ergo*, inferences from the unmodified prior would agree with those from the modified version, provided the simulations of the former did not stray from the relevant region. However, the unmodified posterior would generally have features elsewhere that are wholly incompatible with my prior beliefs. Note that the rationale adopted here ties in closely with that underlying ICM (Besag (1986)), in which a pointwise (local) maximum of the posterior distribution is sought, initiated by maximum likelihood.

Some additional points, again that require further thought, are as follows. The particular modification mentioned above is not sufficient on its own, from the viewpoint of theory rather than practice. If Metropolis method or the Gibbs sampler is run for long enough, with β moderately large, the resulting simulations seem not to be blobby but merely partition the pixels into c essentially single-colour regions. This conflicts with the pictures that are often produced but where sampling has not been continued for a sufficiently long time; this can be extremely long! Thus, I feel that Sigeru Mase's suggestion of "isolated convex-like components" may be overstated, unless there is evidence in the data, though I agree there will be a tendency to smooth out concavity and this may be desirable or undesirable dependent on the context. Incidentally, the fact that simulations of Potts, including Ising, models take so long to reach their proper conclusions adds fuel to the argument that it is their early behaviour, based only on local properties, that is manifest in running the simulations of the posterior. Note here that if the

Swendsen-Wang (1987) algorithm is used for simulation, the above remarks may no longer be relevant, since groups of pixels have their colours changed simultaneously and long-range characteristics can appear more speedily with appreciable probability. Of course, it may occur that the data are in conflict even with the local characteristics of the prior, in which case sampling from the modified and unmodified posteriors will be very different. However, I want to know and to worry when this is the case, so that I can take appropriate action, which might be to reject the data or the prior; this suggests working with the unmodified prior, else one may not notice the conflict. It is worth contrasting tasks in spatial statistics with those in image analysis; in the former, one is not and may never be at the stage of making fully automated decisions. Whatever else, more experience is required before applying simplistic priors as a matter of course.

I come now to the values of hyperparameters. The main problem is with those in the prior, rather than with those in the likelihood. One approach is to preassign the value in the prior, which may well be the best thing to do in a single-parameter family. This can extend to more complicated priors if training data are available. Otherwise, one needs to estimate the hyperparameters. Two methods were suggested in the paper. Yosihiko Ogata feels that the fully Bayesian procedure in Section 4 is the main contribution of the paper. I shall be disappointed if he is correct, since I have the feeling that other methods, based for example on the EM algorithm, may eventually be preferred, despite their failure to allow for hyperparameter uncertainty in constructing interval estimates for the true scene. Rather than discuss this further in the context of mapping, I should like to concentrate here on the example in Section 3 of the paper. This makes use of an entirely *ad hoc* method, borrowed from Besag ((1986), Section 5.1.2), and contravenes one of the basic rules of Bayesian image analysis, namely that one should never lose sight of the data. Briefly, the idea was to obtain a restoration $\hat{x}$ of the image, estimate the likelihood parameters θ from $l(y|\hat{x};\theta)$ and the prior parameters β from $p(\hat{x};\beta)$, then obtain a new restoration, and so on. The problem is that, in estimating β, one loses sight of y. It follows that the starting point of the algorithm is crucial, even when the posterior distribution for fixed β is nicely behaved. For example, if the Potts prior is adopted and the starting point is a single-colour scene, the estimate of β will be infinite, the data will be presumed uninformative and nothing will change. In fact, if maximum pseudo-likelihood estimation is used to estimate β, an infinite value is obtained for any restoration that satisfies the "majority vote" condition; for an example, see Besag ((1986), Fig. 4c), though it should be noted that the true scene was deliberately chosen to have awkward features and, as it happens, the eventual restoration itself is not too bad. Even when things apparently go well, it should be noted that the estimate of β obtained especially from ICM or MAP restorations will tend to be larger than would be found from a typical scene from the posterior distribution.

The above problems can be fixed as follows, using a very simple and neat method introduced by Qian and Titterington (1991). In place of the parameter estimation stage above, one chooses β and θ to maximise the pseudo-likelihood,

$$\prod_{i=1}^{n} P(y_i|\hat{x}_{-i};\beta,\theta).$$

The apparent robustness of this procedure is extremely attractive and is demonstrated on the archaeological data in Figs. R1 to R3. In each case, there is an initiating scene and three subsequent classifications, at stages 1, 4 and 11, each stage consisting of parameter estimation, followed by 1000 cycles of the Gibbs sampler, at the current parameter values, and classification, as in Section 3. Thus, Fig. R1 is initiated by naive classification (cf. the first panel in Fig. 1), Fig. R2 by an array of zeros, and Fig. R3 by an array of ones. There is one very minor modification to the Qian-Titterington algorithm in that the current parameters were chosen to be the medians of the new values and those previously used. This was to prevent alternating between high and low estimates and requires further thought; note that this behaviour does not occur when using ICM, rather than the Gibbs sampler, which was the context of the Qian-Titterington paper. The initial and final estimates of β, for Figs. R1 to R3, were (1.077, 0.768), (0.230, 0.739), (−0.230, 0.737); those for κ were (0.408, 0.435), (0.496, 0.433), (0.496, 0.435). I anticipate that the Qian-Titterington algorithm will become a standard method of parameter estimation in many image analysis problems. It has an underlying rationale and an obvious intuitive appeal. Note that Qian and Titterington (1991) contains more sophisticated versions of the algorithm, though these are possibly of more specialized interest. One point that needs to be borne in mind in programming the algorithm is that, in common with almost any "missing-data" method, concavity cannot be assumed.

This would seem a good point at which to add a comment about more conventional pseudo-likelihood estimation, which is mentioned by Dietrich Stoyan. The idea was introduced in Besag (1975) for realizations from Markov random fields. Rigorous proofs of consistency (mine was not!) appear in Geman and Graffigne (1987), Gidas (1987), Guyon (1987) and Comets (1989). There are still many open problems concerning the efficiency and asymptotic normality of maximum pseudo-likelihood estimators in this context; some results appear in Besag (1977) and Guyon (1987). The technique was extended to Markov point processes in Besag (1978) and there have been several recent papers on this topic, including Jensen and Møller (1989), which establishes consistency. Pseudo-likelihood estimation has also been used for replicate data from complicated exponential family models; for the theoretical basis, see Grenander (1989). In a general setting, pseudo-likelihood techniques have the advantage of simplicity and flexibility; in particular, they avoid the need to evaluate any awkward normalizing constants (partition functions). However, they should not be applied indiscriminately in (e.g. Gaussian) situations where the parameter space is constrained, since the constraints will not necessarily be honoured; nevertheless, see Künsch (1987) for an interesting application to intrinsic Gaussian autoregressions. As computing capabilities continue to increase, one anticipates that pseudo-likelihood methods will gradually be superseded (see, for example, Geyer and Thompson (1991)) but they may retain a place in the type of applications considered in the previous paragraph. The reason is that pseudo-likelihood is concerned essentially with *local* characteristics and hence can be more relevant than methods based, for example, on the full likelihood for the wrong model. It should be borne in mind that, despite the wealth of theoretical and simulation work that exists in the

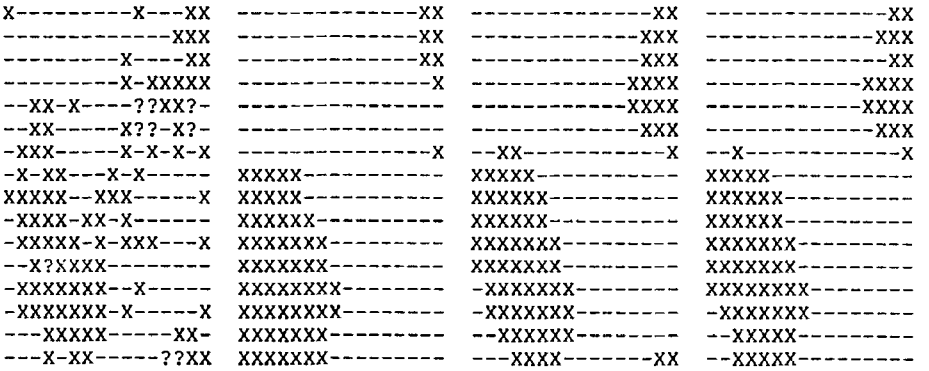

Fig. R1.

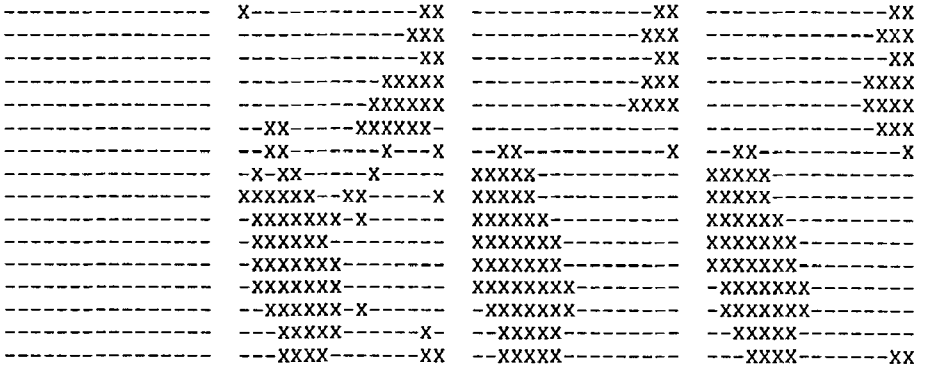

Fig. R2.

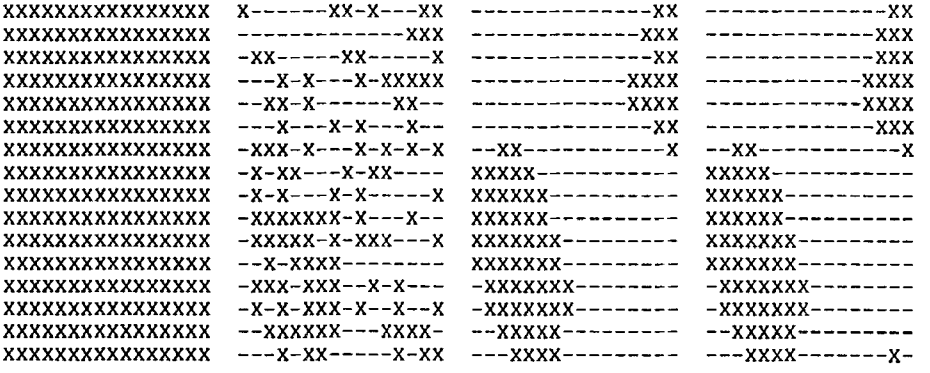

Fig. R3.

Gibbs sampler classifications using Qian-Titterington parameter estimation. In each Figure, the four panels provide the initial scene and classifications after 1000, 4000, and 11000 cycles.

literature, there are very few convincing demonstrations that Markov random fields and especially Markov point processes provide a satisfactory global fit to real data. From this viewpoint, the studies mentioned by Dietrich Stoyan are perhaps of academic rather than practical interest. It should be noted that Diggle *et al.* (1990) limit their investigations to processes with a single interaction parameter, whereas pseudo-likelihood methods extend easily to more complicated situations, particularly since the log pseudo-likelihood for exponential family models is generally concave. Also, the methods can be made reasonably immune to boundary assumptions by appropriate conditioning. Having said all this, I agree with Dietrich Stoyan that, when the model is correct and interaction is strong, pseudo-likelihood is relatively weak, as is shown for Gaussian models in Besag (1977). As regards a means of improving single-pixel pseudo-likelihood, when the model is correct, this can be achieved by looking instead at blocks of pixels. This leads one to wonder whether there is an analogous procedure for point processes.

I should next like to consider some aspects of prior distributions in Bayesian mapping. Recall that the aim in Section 4 was to provide a reasonably accurate and more readily interpretable map of underlying risk within a set of contiguous administrative zones. The data consisted of the number of cases, the number at risk and possibly covariate information for each zone. Ideally, risk should be constant or explained entirely by the covariates, which I take to be the point made by Brian Ripley, though it is difficult to respond to an unpublished example! Otherwise, and this is the usual situation, there remains residual variation and, in the paper, this was thought of as having two possible components, one spatial, the other unstructured, suggesting a corresponding prior distribution in the form of a convolution. Several discussants commented particularly on the form of the spatial component, so let me begin there.

If one is to use a Gaussian prior, then there are two main approaches. The first is that in which one models the covariance structure by a positive definite or semi-definite matrix, as in the geostatistical or "kriging" approach, outlined by Adrian Raftery and Jeffrey Banfield. Note that the inclusion of a "nugget" effect is already catered for by the non-spatial component v. The geostatistical approach has much to offer and clearly provides a worthwhile line of future research in epidemiology, especially since it does not require stationarity and treats areas as areas, rather than as points. In principle, I am less attracted by its usual reliance on distance as a suitable metric but this is probably unimportant at the required level of approximation.

The alternative, pursued in the paper, is to "model" the precision structure of the spatial component. This approach is really the dual of that above, as is implicit in Adrian Raftery and Jeffrey Banfield's discussion. Note that they concentrate for definiteness on the case of finite variances but that the extension to intrinsic processes is immediate – and necessary in the present context. They also express concern about the choice of neighbourhood system, especially when the zones have very irregular sizes and shapes. I am somewhat embarrassed that I have no experiments to report, particularly since they could be easily carried out; certainly they will happen in the near future. The most obvious thought is to include adjacencies of adjacencies as neighbours. Donald and Stuart Geman

raise a more interesting idea in possibly adapting the neighbourhood structure to external factors by some objective rule. They also comment on the closely related issue of the rather strange conditional variance structure in equation (4.3). Lest anyone think that this is a deliberate choice, let me describe how it comes about. It may be helpful to any readers outside the field to begin by considering the salient features of conditional autoregressions, partly because the standard book reference, Ripley ((1981), p.88), assumes invariant conditional variances, which is of little direct relevance to irregularly distributed sites and, even for regular arrays, is questionable at the boundary.

A conditional Gaussian autoregression or auto-Normal scheme (Besag (1974, 1975)), for a random vector $u = (u_1, u_2, \ldots, u_n)^T$, has conditional moment structure,

$$\text{(R1)} \qquad E(u_i|u_{-i}) = \sum_{j \neq i} \beta_{ij} u_j,$$

$$\text{(R2)} \qquad \text{Var}(u_i|u_{-i}) = \kappa_i > 0,$$

where the parameters are subject to further constraints, described below; any mean structure has been omitted here without loss. If $\beta_{ij} = 0$, u_i and u_j are conditionally independent; otherwise i and j are neighbours. Write B for the matrix with (i, j) element,

$$B_{ij} = \begin{cases} 1, & i = j, \\ -\beta_{ij}, & i \neq j. \end{cases}$$

Provided B is non-singular, it follows that $E(u) = 0$. Also u has dispersion matrix $V = B^{-1}\Lambda$ where $\Lambda = \text{diag}\{\kappa_i\}$. The additional conditions for a proper Gaussian distribution are now seen to be that $\Lambda^{-1}B$ must be symmetric and positive definite, the former condition being equivalent to

$$\beta_{ij}\kappa_j = \beta_{ji}\kappa_i, \quad i \neq j.$$

In fact, it can easily be shown that $\text{sign}(\beta_{ij})\sqrt{(\beta_{ij}\beta_{ji})}$ is the partial correlation coefficient between u_i and u_j for $i \neq j$. It is clear that any zero-mean Gaussian distribution can be formulated in the above manner, so the only novelty is that B and Λ are to be specified, rather than dealing immediately in terms of V itself. Various properties can be deduced, including the fact that the right-hand side of (R1) is the best linear predictor of u_i, given u_{-i}.

Having once decided on the neighbourhood ∂i of each site i, it remains to specify the non-zero parameters. It can be assumed that the graph induced by the ∂i's is connected, else the system can be broken down into two or more independent subsystems. The simplest choice is for the conditional mean in (R1) to be proportional to the observed mean $\bar{u}_i$ of the n_i neighbour values. Thus,

$$\text{(R3)} \qquad E(u_i|u_{-i}) = \gamma\bar{u}_i,$$

$$\text{(R4)} \qquad \text{Var}(u_i|u_{-i}) = \kappa/n_i,$$

where $\kappa > 0$ and γ remain to be specified or estimated. There are two points to be noted here. The first is that this defines a valid auto-Normal scheme for any $\gamma \in (-1, 1)$. The second is that the form of the conditional variance (R4) is forced by (R3). A more general formulation is

$$\text{(R5)} \qquad E(u_i|u_{-i}) = \gamma \sum_{j \in \partial_i} w_{ij} u_j / w_{i+},$$

$$\text{(R6)} \qquad \mathrm{Var}(u_i|u_{-i}) = \kappa / w_{i+},$$

in which the w_{ij}'s are symmetric weights and $+$ denotes summation over the corresponding subscript. It does not seem possible to fix the marginal variances of the u_i's to be equal but note that

$$E\left[w_{i+}\left(u_i - u_i^*\right)^2\right] = \kappa,$$

where u_i^* is the best linear unbiased predictor of u_i on the right-hand side of (R5).

In Clayton and Kaldor (1987) and Mollié and Richardson (1991), $\bar{u}_i$ in (R3) is replaced by the sum of the n_i neighbour values. This leads to an invariant *conditional* variance, though the marginal variances can be entirely disparate. Not only does there seem no good reason to base the conditional expectation in (R3) on a sum rather than a mean, when sites have differing numbers of neighbours, but also the maximum value $\gamma_{\max}$ of γ is a function of the neighbourhood structure; thus, $\gamma_{\max}$ can be influenced by changes in the neighbourhood structure in a remote portion of the graph.

The unattractiveness of (R4) is especially acute when γ is very small or zero. It is partly for this reason that we introduced the two-component $u + v$ formulation in Section 4, where u corresponds to (R3) and (R4), with $\gamma = 1$, and v represents independent white noise with variance λ. A further reason is that even moderate spatial dependence requires a value of γ close to unity, as is demonstrated for regular arrays in Besag (1981). Thus, the use of an (infinite variance) intrinsic autoregression in tandem with v provides a means of representing spatial and non-spatial components and, as a bonus, dispenses with the estimation of γ itself. A perhaps naive interpretation of (R4) is that, as an approximation, the greater the number of nearby locations at which one knows the true risk, the more precise is one's prediction at a central site. Whether one can devise some sort of connection between a continuum process, such as a Brownian sheet, and an intrinsic autoregression is an open question, so far as I am aware; any approximations or alternative formulations would be of considerable interest. Note that the basic intrinsic autoregression can be generalized to (R5) and (R6) with $\gamma = 1$ and that this can be used to take account of the features of contiguous zones, as is mentioned in Section 4. In fact, this extended formulation is just (4.1) with $\phi(z) = z^2/2\kappa$, so there is nothing new here. One special case that might be of interest is $w_{ij} = m_i m_j$, where m_i is a property of site or zone i. Then

$$E(u_i|u_{-i}) = \sum_{j \in \partial i} m_j u_j / m_i^+, \quad \mathrm{Var}(u_i|u_{-i}) = \kappa / m_i m_i^+,$$

where

$$m_i^+ = \sum_{j \in \partial i} m_j.$$

In the mapping context, m_i might be the population at risk in zone i, for example.

As in earlier discussion of Potts priors, there is no attempt above to do any serious modelling, though the separation into extra Poisson (v) and spatial (u) components has some credibility. However, one might go a little further with the intrinsic component, though I shall only discuss this in the context of fitting a surface over a square array of sites at each of which a noise-degraded value is available. The neighbourhood criterion will be assumed to be translation invariant. In the simplest case, with equal weight ascribed to each neighbour, the predicted value at any site, given all other values, is that obtained by fitting a plane to the neighbour values by ordinary least squares. Instead of this, one might choose to fit a 6-parameter quadratic surface to these values. For example, with an 8-site second-order neighbourhood, directly and diagonally adjacent sites then receive respective weights $1/2$ and $-1/4$. In fact, this intrinsic autoregression can be shown to be degenerate, in that its generalized spectrum is the product of two one-dimensional spectra; that is, the process is *separable*. This is not the case for the corresponding 12-site third-order autoregression, which assigns weights $1/4$, $1/8$ and $-1/8$ to first-, second- and third-order neighbour values, respectively. The aim of using such a prior would be to encourage locally quadratic rather than locally planar behaviour in realizations from the posterior distribution. Incidentally, in representing *textures* by auto-Normal schemes, the parameter values seem invariably to be extremely close to the boundary of stationarity and one wonders whether such models might profitably be replaced by intrinsic autoregressions.

Archaeological data

I am grateful to Adrian Raftery and Jeffrey Banfield for producing an alternative analysis of the soil phosphate data and I am relieved that our classifications are very similar. As we stated in the paper, there are many different ways in which such classifications could be produced and I think it is fair to say that here the task is not particularly difficult. This is in complete contrast with the very much harder problem, concerning the identification of ice floes from synthetic aperture radar data, that is tackled with great success in Banfield and Raftery (1989). In a sense, it is almost detrimental to their work that they bother to tackle the archaeological example!

As regards our analysis, the main aim was really to provide a probabilistic classification from the data. I am not sure as to the "precise status of this statement of uncertainty" but, in this instance, I think I would stand by Fig. 2 as a reasonable representation of posterior probability and, were I to provide the archaeologist with some results, I would certainly rather produce Fig. 2 than a unique classification. Although I don't wish to pin too much faith on a single example, particularly given its circumstances, I am optimistic that further research and modifications will lead to wider applications. As a postscript to the example, I have just noticed that the final panel of Fig. 1, which we copied manually from

APL character output, contains an extra blob that I hope will be fixed in the final production. If not, it is not too serious; the associated posterior probability in Fig. 2 is 0.45!

Epidemiology data

The data sets in Section 4 of the paper were intended to illustrate different aspects of Bayesian mapping. The French data are based on large zones (départements), that are fairly regular in shape (I am indebted to Adrian Raftery for the reason!). Such data should not present too many problems, though the assumption of constant κ and λ across the country is perhaps a weakness. Two particular examples from a larger set were chosen: thyroid cancer to display appreciable spatial structure and multiple myeloma to display its absence. Thus, the inclusion of an example for which κ is very small, as indeed is λ, was deliberate. Several discussants comment on these small values and point out that the choice of ε is critical. No doubt this is true numerically but my conclusion was that the values are sufficiently small to indicate almost constant risk; compare Figs. 9 and 10. The reason for excluding the origin is not because one refuses to believe that $\kappa = \lambda = 0$ is possible but is purely computational. I can assure Peter Green that no experimentation was done before we adopted (4.6), though this would be a good idea in a more general context. However, I return to my earlier comment that methods of estimating κ and λ, other than the Gibbs sampler, may well prevail. In a sense, this would turn full circle. In Clayton and Kaldor (1987), the EM algorithm was adopted for estimating hyperparameters. Subsequently, David Clayton suggested the Gibbs sampler for this purpose, so far as I can recall, and this was the method we adopted. Of course, the Gibbs sampler may survive, perhaps using a version of the suggestion by Adrian Raftery and Jeffrey Banfield. Scale invariance would certainly have an advantage. The third example, concerning the incidence of solid tumours in Greater Manchester, was included because it involved a rather small number of cases among the 216 zones (electoral wards) and produced an interesting result that was not obvious at first sight in the raw data (Fig. 14). As we mentioned, the data are part of a larger data set for 1218 wards in the North of England. I should emphasize that the reason for concentrating on Manchester alone was not computational but a matter of presentation. Even with my somewhat pedestrian programming in APL, I am able to run the Gibbs sampler on all 1218 wards simultaneously. I agree entirely with Peter Green that such methods can already be implemented in image analysis, *per se*, at least for problems of modest size, though not routinely in the sense of real-time processing. In fact, I would go further in suggesting that, if one has a good solution to a class of real problems and if one can "sell" it (this is perhaps the hardest part!), then appropriate hardware will be devised. I think the benefit of small problems, and this includes looking at interesting parts of larger images, is that one can experiment more easily, though Brian Ripley questions the relevance of this; I think we shall have to agree to disagree here.

Returning to the analysis of the Manchester data, I am impressed by the endeavour of Yosihiko Ogata in illustrating his own approach, though I am not sure I agree with the results. In his original analysis, he worked from the relative

incidence rates in Fig. 14 but, of course, was not able to allow for the important fact that these were calculated using denominators that differed by a factor of almost four and hence gave rise to considerable differences in precision. The use of a Gaussian distribution with unconstrained parameters provides an alternative to the inclusion of v. However, as we stated, the idea was to avoid a Gaussian model because of the low numbers. This admittedly becomes much more important in rural areas: 165 of the 1218 wards contain no case and 183 only one, whereas the minimum in Greater Manchester is two. For his later analysis, Yosihiko Ogata used a Poisson likelihood but still omitted v from the prior. I think one needs to be very careful here. For example, in running one of Jeremy York's programs on the remainder of the Manchester cancer data, that is for leukaemias, I apparently discovered a fairly strong pattern, which one would not expect. Eventually, I realized that I had inadvertently used a version that excluded v and that therefore any extra-Poisson variation was being forced on u. When the full version was run, the pattern disappeared. I doubt that this phenomenon is rare. Incidentally, this suggests that extra-Poisson variation can be teased out and I think this has been quite a common assumption in biostatistics since the pioneering work of Breslow (1984). Yosihiko Ogata's Fig. F2 (but surely not Fig. F4?) is indeed very close to our Fig. 15, though his $\hat{\tau}^2$ is somewhat larger than our $\hat{\kappa} = 0.024$. However, our main difference is that I would certainly reject constant risk as a rival explanation. Incidentally, if constant, the overall relative risk is unity by design, rather than 1.07, which is the mean of the relative incidence rates and does not allow for differing numbers at risk. As was stated in the paper, the same residential/industrial effect occurs in the other major industrial conurbation, Tyne and Wear. An effect such as this is likely to be small, otherwise it would already be well known. My personal view is that one should investigate whether such an effect has a reasonable explanation and whether it is repeated elsewhere, rather than rely on somewhat arbitrary statistical tests; see our comments on the example.

Yosihiko Ogata is not alone in commenting on the scant attention we paid to model uncertainty; the point is raised by several discussants. Whilst I accept the criticisms in general, I find the idea of attaching prior probabilities to different models as too subjective a matter in the present context! I admit that the use of L1 as well as L2 priors (equations (4.4) and (4.2)) in our examples was primarily to illustrate that other possibilities are open. Peter Green's discussion of "proposals" is very useful in this respect. We are already indebted to him for advice on carrying out the Gibbs sampler simulations in Section 4. Finally, I had thought that the simulation exercise based on the estimated risk surface for the thyroid cancer data provided a reasonably rigorous means of comparing different methods and models, though I admit that our only comparison was with the raw incidence rates in the simulated data, which can hardly be described as stringent. However, to put the ball back in Donald and Stuart Geman's court, what elementary methods would they suggest for dealing with such data? Recall that zones are generally irregular in shape and that the incidence rates are very low and/or the populations at risk are small. Judging from typical cancer atlases, satisfactory methods are not widely available.

Running the Gibbs sampler

Adrian Raftery and Jeffrey Banfield have tackled the difficult problem of sampling strategy in running the Gibbs sampler. I agree that discarding realizations is wasteful in principle. In the epidemiology examples, we did this merely to avoid storage problems, which become a little more acute for all 1218 wards, though this is really a minor issue. No realizations were discarded in the archaeological example, incidentally. I think the ideas behind Adrian Raftery and Jeffrey Banfield's calculations have great potential but I question whether we could reliably have used far less simulation. The problem is that the Gibbs sampler can get stuck near $\kappa = 0$ or $\lambda = 0$ for long periods, despite the $\varepsilon-$fix; this gets worse as ε is decreased, explaining my reluctance to use a value smaller than 0.01, though valid in theory. Two possible remedies are (i) use some other method of estimating κ and λ or (ii) amend the Gibbs sampler transition probabilities in some allowable fashion, so as to exit from small values of κ and λ more easily. Note that the problem exists even when the posterior means of κ and λ are relatively large.

And finally ...

There are many issues that have not been dealt with satisfactorily in my response and others that have not even been touched on. For example, Adrian Raftery and Jeffrey Banfield raise the possibility of using conditional probabilities in the prior that do not generate a valid Markov random field because of their mutual incompatibility. I have no objection in principle to such devices which can be thought of as approximating genuine distributions, provided the conditional probabilities are not too bizarre. However, as Donald and Stuart Geman point out, there are already many somewhat arbitrary decisions to be made and it may be prudent not to add more.

Once again, I am very grateful to the discussants for their insightful comments and hope that the paper, as a whole, will stimulate future research in spatial statistics.

Additional References

Amit, Y., Grenander, U. and Piccioni, M. (1991). Structured image restoration through deformable templates, *J. Amer. Statist. Assoc.* (to appear).

Besag, J. E. (1977). Errors-in-variables estimation for Gaussian lattice schemes, *J. Roy. Statist. Soc. Ser. B*, **39**, 73–78.

Besag, J. E. (1978). Some methods of statistical analysis for spatial data (with discussion), *Bull. Internat. Statist. Inst.*, **47**, Book 2, 77–92.

Besag, J. E. (1981). On a system of two-dimensional recurrence equations, *J. Roy. Statist. Soc. Ser. B*, **43**, 302–309.

Besag, J. E. and Clifford, P. (1989). Generalized Monte Carlo significance tests, *Biometrika*, **76**, 633–642.

Besag, J. E. and Clifford, P. (1991). Sequential Monte Carlo p-values, *Biometrika*, **78** (to appear).

Comets, F. (1989). On the consistency of a class of estimators for exponential families of Markov random fields on the lattice, Universite de Paris–X (preprint).

Fredkin, D. R. and Rice, J. A. (1991). Bayesian restoration of single channel patch clamp recordings, *Biometrics* (to appear).

Geman, D. (1991). Random fields and inverse problems in imaging, *Lecture Notes in Math.*, Springer, Berlin (to appear).

Geman, D. and Reynolds, G. (1991). Constrained restoration and the recovery of discontinuities, Tech. Report, Department of Mathematics and Statistics, University of Massachusetts at Amherst.

Geyer, C. J. and Thompson, E. A. (1991). Maximum likelihood in exponential families (with discussion), *J. Roy. Statist. Soc. Ser. B*, **53** (to appear).

Gidas, B. (1987). Consistency of maximum likelihood and pseudolikelihood estimators for Gibbs distributions, *Proceedings of the Workshop on Stochastic Differential Systems with Applications*, Springer.

Green, P. J. (1986). Discussion: On the statistical analysis of dirty pictures, by Besag, J. E., *J. Roy. Statist. Soc. Ser. B*, **48**, 259–302.

Grenander, U. (1989). Advances in pattern theory, *Ann. Statist.*, **17**, 1–30.

Guyon, X. (1987). Estimation d'un champ par pseudo-vraisemblance conditionelle: étude asymptotique et application au cas markovien, *Actes de la 6éme Rencontre Franco-Belge de Statisticiens*, Edition des Facultés Universitaires de Saint-Louis, Bruxelles.

Jensen, J. L. and Møller, J. (1989). Pseudolikelihood for exponential family models of spatial processes, Research Report, No. 203, Department of Theoretical Statistics, Institute of Mathematics, University of Aarhus, Denmark.

Lippman, A. (1991). Markov models for speech recognition, *Spatial Statistics and Image Processing: a Cross-Disciplinary Report* (eds. J. J. Simpson and J. E. Besag), National Academy of Sciences Press, Washington, D. C.

Mollié, A. and Richardson, S. (1991). Empirical Bayes estimates of cancer mortality rates using spatial models, *Statist. Med.*, **10**, 95–112.

Qian, W. and Titterington, D. M. (1991). Stochastic relaxations and E-M algorithm for Markov random fields, *J. Statist. Comput. Simulation* (to appear).

Ripley, B. D. (1981). *Spatial Statistics*, Wiley, New York.

Sheehan, N. A. (1990). Image processing procedures applied to the estimation of genotypes on pedigrees, Tech. Report, No. 176, Department of Statistics, University of Washington.

Swendsen, R. H. and Wang, J-S. (1987). Non-universal critical dynamics in Monte Carlo simulations, *Phys. Rev. Lett.*, **58**, 86–88.

Thomas, A. (1991). Genotypic inference with the Gibbs sampler, *Proceedings for the Workshop on Analytic Methods for Population Viability and Analyses, Front Royal, Virginia, October 1989* (to appear).