Julian Besag FRS, 1945–2010

Peter Green

School of Mathematics
University of Bristol

23 August 2011 / ISI World Statistics Congress

Outline

1. Biography
2. Methodology
3. Applications
4. Last seminars

Julian Besag
as a young boy

A brief biography

1945	Born in Loughborough
1965–68	BSc Mathematical Statistics, Birmingham
1968–69	Research Assistant to Maurice Bartlett, Oxford
1970–75	Lecturer in Statistics, Liverpool
1975–89	Reader (from 1985, Professor), Durham
1989–90	Visiting professor, U Washington
1990–91	Professor, Newcastle-upon-Tyne
1991–2007	Professor, U Washington
2007–09	Visiting professor, Bath
2010	Died in Bristol

Visiting appointments in Oxford, Princeton, Western Australia, ISI New Delhi, PBI Cambridge, Carnegie-Mellon, Stanford, Newcastle-u-Tyne, Washington, CWI Amsterdam, Bristol

Julian Besag in 1976

Methodology

- Spatial statistics
  - Modelling, conditional formulations
  - Frequentist and Bayesian inference
  - Algebra of interacting systems
  - Digital image analysis
- Monte Carlo computation and hypothesis testing
- Markov chain Monte Carlo methods
- Exploratory data analysis

Modelling, conditional formulations: key papers

Nearest-neighbour systems and the auto-logistic model for binary data. *Journal of the Royal Statistical Society B* (1972).

Spatial interaction and the statistical analysis of lattice systems (with Discussion). *Journal of the Royal Statistical Society B* (1974).

On spatial-temporal models and Markov fields. *Proceedings of the 10th (1974) European Meeting of Statisticians* (1977).

The 1974 paper in *JRSS(B)*

Spatial Interaction and the Statistical Analysis of Lattice Systems

By Julian Besag

University of Liverpool

[Read before the Royal Statistical Society at a meeting organized by the Research Section on Wednesday, March 13th, 1974, Professor J. Durbin in the Chair]

Summary

The formulation of conditional probability models for finite systems of spatially interacting random variables is examined. A simple alternative proof of the Hammersley–Clifford theorem is presented and the theorem is then used to construct specific spatial schemes on and off the lattice. Particular emphasis is placed upon practical applications of the models in plant ecology when the variates are binary or Gaussian. Some aspects of infinite lattice Gaussian processes are discussed. Methods of statistical analysis for lattice schemes are proposed, including a very flexible coding technique. The methods are illustrated by two numerical examples. It is maintained throughout that the conditional probability approach to the specification and analysis of spatial interaction is more attractive than the alternative joint probability approach.

Keywords: Markov fields; spatial interaction; auto-models; nearest-neighbour schemes; statistical analysis of lattice schemes; coding techniques; simultaneous bilateral autoregressions; conditional probability models

JRSS(B) 1974: the Hammersley–Clifford theorem

The theorem states that "Markov random fields are the same as Gibbs distributions" – that is, that a multivariate distribution satisfies the Markov random field property if and only if its log-density is additive over cliques.
Besag gave a simple proof of this, assuming "positivity" – essentially that any combination of values realisable locally was realisable globally.

JRSS(B) 1974: the Hammersley–Clifford theorem

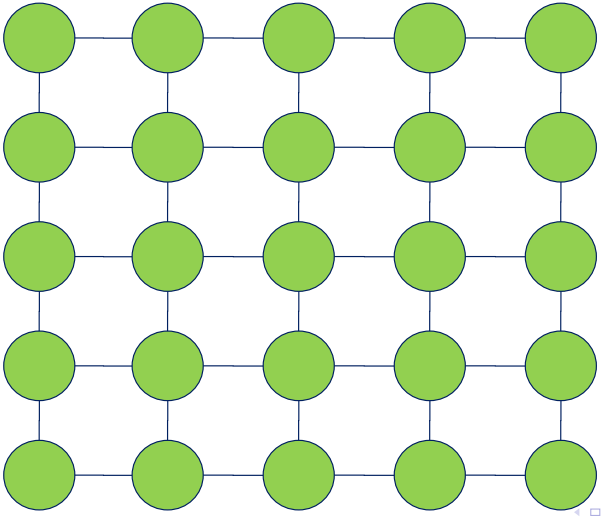
Proof of theorem. It follows from equation (3.2) that, for any $\mathbf{x} \in \Omega$, $Q(\mathbf{x}) - Q(\mathbf{x}_i)$ can only depend upon x_i itself and the values at sites which are neighbours of site i. Without loss of generality, we shall only consider site 1 in detail. We then have, from equation (3.3),

$$Q(\mathbf{x}) - Q(\mathbf{x}_1) = x_1 \Big\{ G_1(x_1) + \sum_{2 \leqslant j \leqslant n} x_j G_{1,j}(x_1, x_j) + \sum\sum_{2 \leqslant j < k \leqslant n} x_j x_k G_{1,j,k}(x_1, x_j, x_k) + \dots + x_2 x_3 \dots x_n G_{1,2,\dots,n}(x_1, x_2, \dots, x_n) \Big\}.$$

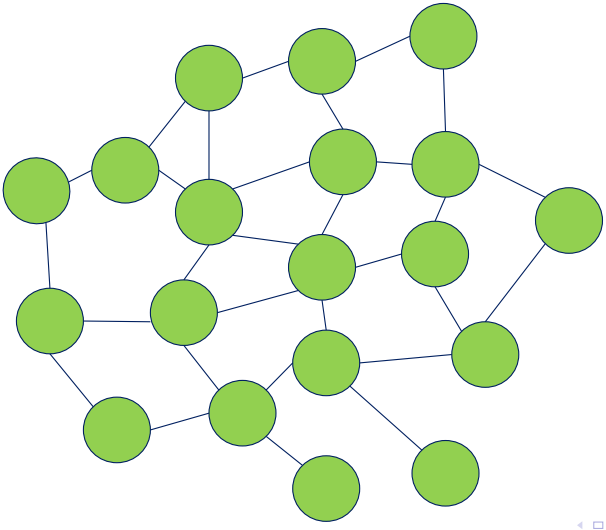
Now suppose site $l\,(\neq 1)$ is *not* a neighbour of site 1. Then $Q(\mathbf{x}) - Q(\mathbf{x}_1)$ must be independent of x_l for all $\mathbf{x} \in \Omega$. Putting $x_i = 0$ for $i \neq 1$ or l, we immediately see that $G_{1,l}(x_1, x_l) = 0$ on Ω. Similarly, by other suitable choices of $\mathbf{x}$, it is easily seen successively that all 3-, 4-, …, n-variable G-functions involving both x_1 and x_l must be null. The analogous result holds for any pair of sites which are not neighbours of each other and hence, in general, $G_{i,j,\dots,s}$ can only be non-null if the sites $i, j, \dots, s$ form a clique.

On the other hand, any set of G-functions gives rise to a valid probability distribution $P(\mathbf{x})$ which satisfies the positivity condition. Also since $Q(\mathbf{x}) - Q(\mathbf{x}_i)$ depends only upon x_l if there is a non-null G-function involving both x_i and x_l, it follows that the same is true of $P(x_i \mid x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n)$. This completes the proof.

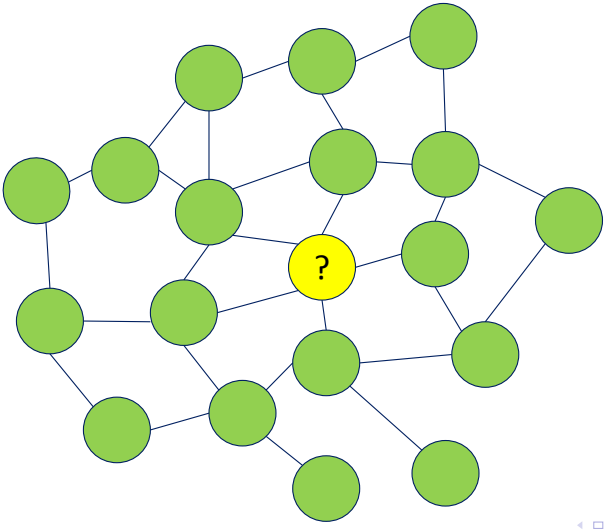
Markov random fields



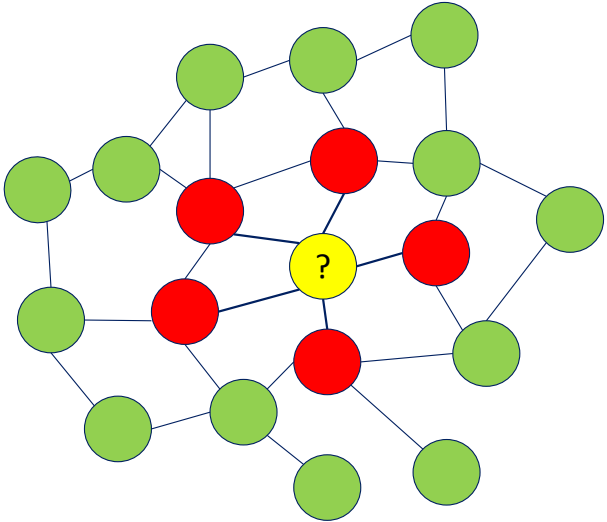
Markov random fields



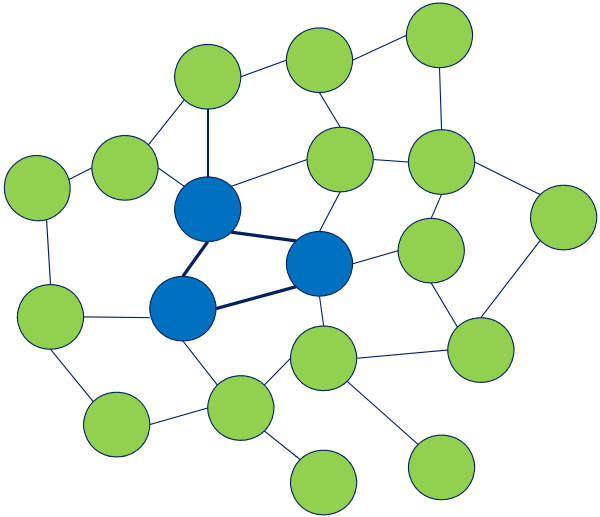
Markov random fields



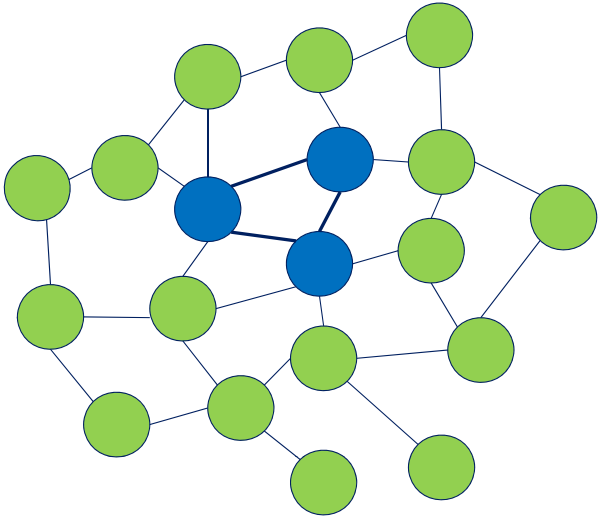
Markov random fields



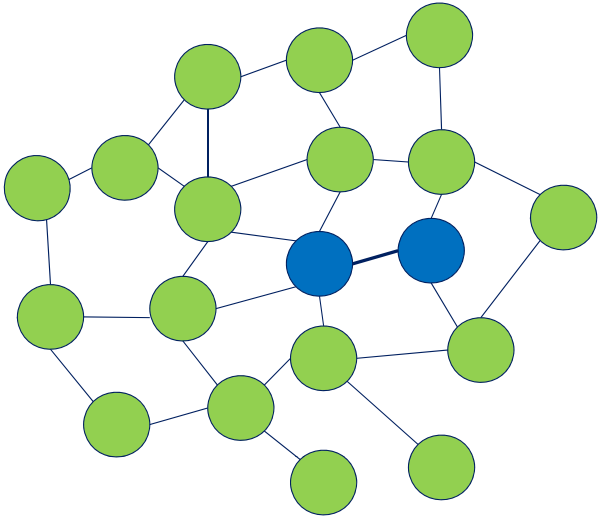
Markov random fields = Gibbs distributions



Markov random fields = Gibbs distributions



Markov random fields = Gibbs distributions



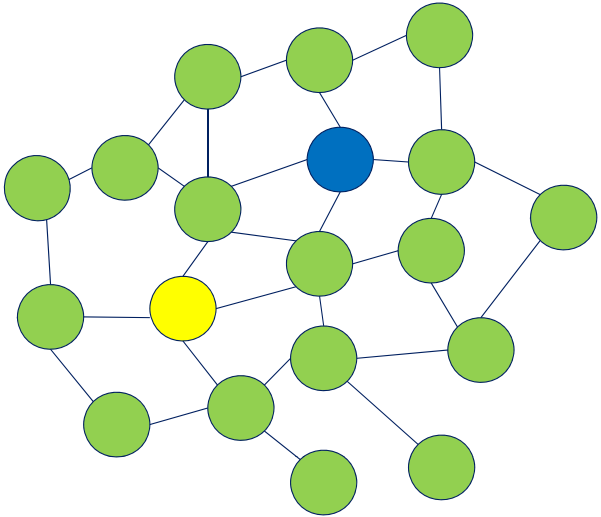
The Hammersley–Clifford theorem

Many years later, the theorem was superseded by a more complete understanding of Markov properties in undirected graphical models: we can distinguish Global, Local and Pairwise Markov properties, and relate all these to the Factorisation property of Gibbs distributions; in general

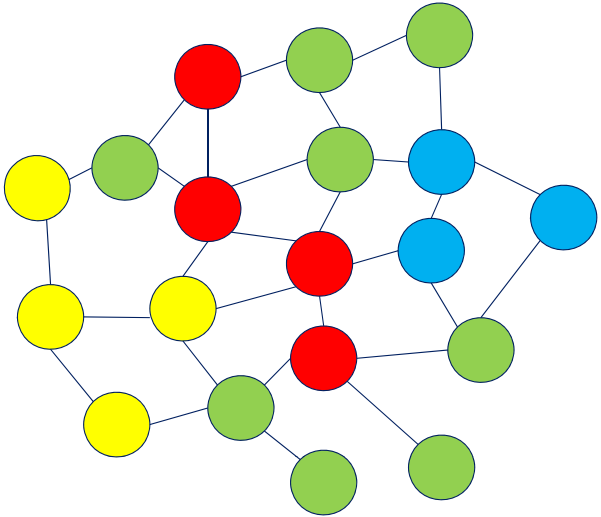
$$F \implies G \implies L \implies P$$

and under an additional condition implied by positivity they are all equivalent.

Pairwise Markov property



Global Markov property



Other key ideas in the 1974 paper

- Auto-models
  - auto-normal, auto-logistic, etc
  - establishes idea that the natural multivariate generalisations of standard univariate distributions might be based on conditional not marginal distributions having the assumed form.
- Inferential methods for MRFs
- Ecological applications
- Characterisation of MRFs as equilibria of certain space-time models

Inference in spatial systems: key papers

Statistical analysis of non-lattice data. *The Statistician* (1975).

On the estimation and testing of spatial interaction in Gaussian lattice processes. *Biometrika* (1975). (with Pat Moran)

Errors-in-variables estimation for Gaussian lattice schemes. *Journal of the Royal Statistical Society B* (1977).

Efficiency of pseudo-likelihood estimation for simple Gaussian fields. *Biometrika* (1977).

Pseudo-likelihood estimation

3.3. *A Pseudo-likelihood Technique*

Given the previous set-up, perhaps the most naive approach to the estimation of the unknown parameters in the terms $p_i(\psi)$ would be to take that vector $\tilde{\psi}$ which maximizes the quantity

$$L_n(\tilde{\psi}) = \sum_{i=1}^{n} \ln p_i(\psi) \qquad (3.2)$$

with respect to ψ. Of course, L_n is not the true log-likelihood function for the sample (except in the trivial case of complete independence) and yet its maximization, especially in view of the coding technique, would seem to present an intuitively plausible method of estimation. That this intuition can be given a theoretical foundation will be seen later on. Note

Pseudo-likelihood estimation

Thus in a Markov random field x, indexed by sites $i \in S$, with parameter ψ, the maximum pseudo-likelihood estimator of ψ is that maximising

$$\prod_{i \in S} p(x_i | x_{S \setminus i}, \psi)$$

... evidently not the joint probability of anything! It is motivated by Besag's previous 'coding technique' estimator where the product is restricted to sites i that are not neighbours of each other, when it is a straightforward conditional likelihood.
It has proved a powerful and durable idea in various contexts with dependent data, notching up now 733 citations.

Pseudo-likelihood estimation

We now return to the general problem of providing some mathematical justification for maximum pseudo-likelihood estimators. The only property which will be established is that of consistency and, as such, we shall have to admit a conceptual passage of n to the infinite limit. How relevant this is to spatial applications, where n is usually fixed, is a matter for debate; for example, the imagination palls at the thought of extending the counties of Eire to an infinite set! Nevertheless, the property of consistency might be thought of as a minimal statistical requirement. We shall sketch its proof; a rigorous treatment would, for example, require some consideration of the system boundary as n increases.

Algebra of interacting systems: key papers

On a system of two-dimensional recurrence relations. *Journal of the Royal Statistical Society B* (1981).

On conditional and intrinsic autoregressions. *Biometrika* (1995) (with Charles Kooperberg).

Markov random fields with higher-order interactions. *Scandinavian Journal of Statistics* (1998) (with Håkon Tjelmeland).

A recursive algorithm for Markov random fields. *Biometrika* (2002) (with Francesco Bartolucci).

First-order intrinsic autoregressions and the de Wijs process. *Biometrika*, 92, 909-920 (2005) (with Debashis Mondal).

Intrinsic autoregression and the de Wijs process

First-order intrinsic autoregressions and the de Wijs process

BY JULIAN BESAG AND DEBASHIS MONDAL

Let $\{X_{u,v}: u, v \in \mathscr{Z} = 0, \pm 1, \dots\}$ be a homogeneous first-order intrinsic autoregression on the two-dimensional rectangular lattice $\mathscr{Z}^2$ (Künsch, 1987; Besag & Kooperberg, 1995; Rue & Held, 2005, Ch. 3), with generalised spectral density function

$$f(\omega, \eta) = \kappa(1 - 2\beta \cos \omega - 2\gamma \cos \eta)^{-1} \quad (\omega, \eta \in (-\pi, \pi]) \tag{1}$$

and conditional expectation structure

$$E(X_{u,v} | \dots) = \beta(x_{u-1,v} + x_{u+1,v}) + \gamma(x_{u,v-1} + x_{u,v+1}), \quad \text{var}(X_{u,v} | \dots) = \kappa, \tag{2}$$

where $\beta, \gamma > 0$ and $\beta + \gamma = \frac{1}{2}$. Note that we use the term 'order' to identify the neighbour-

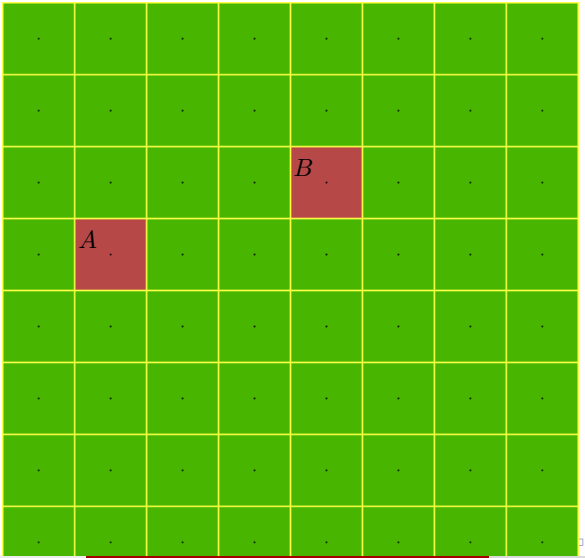
Intrinsic autoregression and the de Wijs process

We close this section with some brief remarks about the de Wijs process (de Wijs, 1951, 1953; Matheron, 1971), a generalised continuum formulation that takes nonzero values with respect to contrasts between averages over areas rather than values at points. The restriction to averages is relevant in practice, where 'point' measurements are generally idealisations. The de Wijs process is Gaussian and Markovian and its variogram intensity increases as the logarithm of distance, which underlies its invariance to conformal transformations and its physical appeal as a basic model. To be more specific, the de Wijs process $\{Y_\varphi\}$ is a generalised Gaussian Markov process indexed by functions φ on the plane that integrate to zero and that give rise to the well-defined variance formula

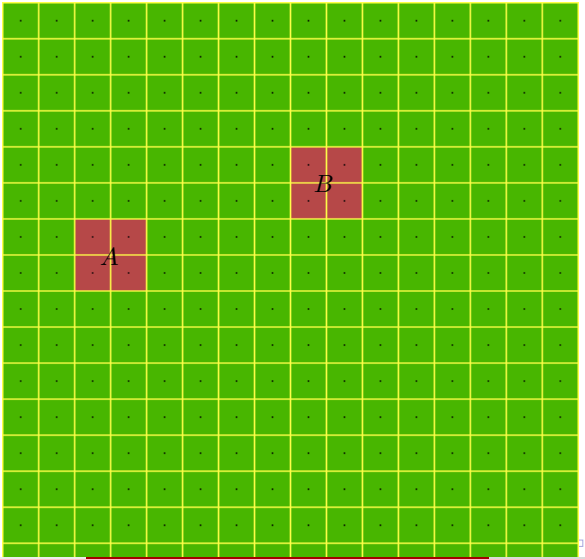
$$\operatorname{var}(Y_\varphi) = -\int\int \log\|x-y\|\varphi(x)\varphi(y)\,dx\,dy.$$

In particular, let $\varphi(x) = 1_A(x)/|A| - 1_B(x)/|B|$, where A and B are two regions of the plane with respective areas $|A|>0$ and $|B|>0$. Then we interpret Y_φ as the difference in the average values Y_A and Y_B of the de Wijs process over A and B. These averages correspond to an intrinsic process with a single degeneracy and generalised variogram defined by $v_{A,B} = \frac{1}{2}\operatorname{var}(Y_A - Y_B)$. Note that the spectral density function of the de Wijs process is inversely proportional to $\omega^2+\eta^2$. We return to this in § 4·2 together with some generalisations.

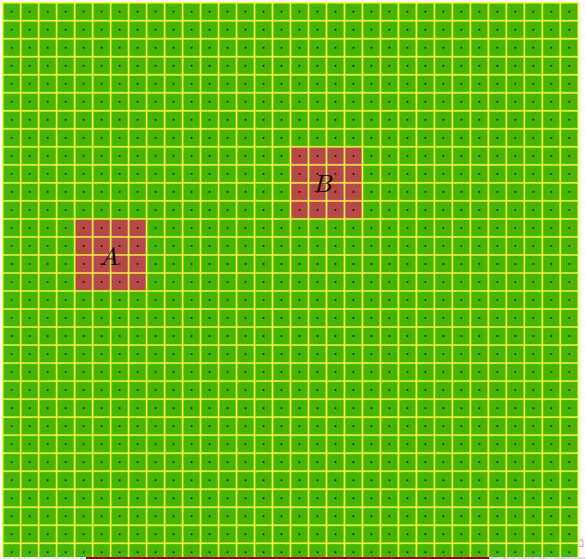
Intrinsic autoregression and the de Wijs process



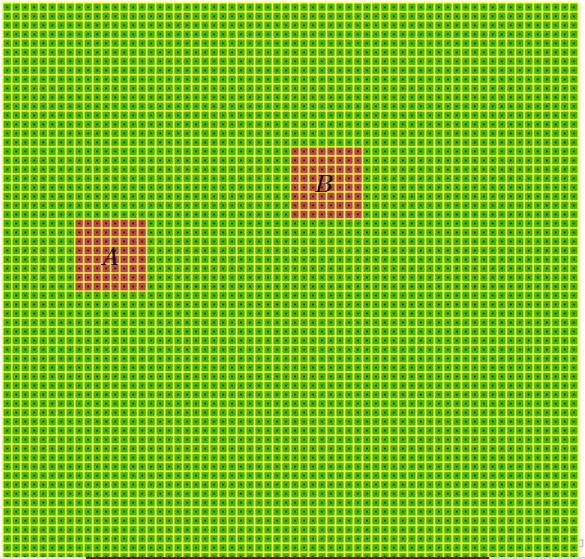
Intrinsic autoregression and the de Wijs process



Intrinsic autoregression and the de Wijs process

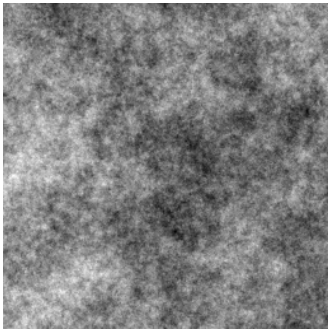


Intrinsic autoregression and the de Wijs process

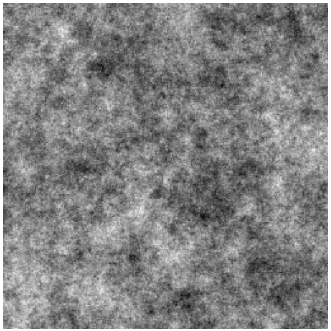


Intrinsic autoregression and the de Wijs process

Integrated de Wijs process



Intrinsic autoregression



... provides a rigorous link between geostatistical models and (intrinsic) lattice Markov random fields, explaining empirical robustness of discrete-space formulations to changes of scale.

Digital image analysis: key papers

On the statistical analysis of dirty pictures (with Discussion). *Journal of the Royal Statistical Society B* (1986).

Towards Bayesian image analysis. *Journal of Applied Statistics* (1989).

Bayesian image restoration, with two applications in spatial statistics (with Discussion). *Annals of the Institute of Statistical Mathematics* (1991) (with Jeremy York and Annie Mollié)

The 1986 paper in *JRSS(B)*

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.

Dirty pictures

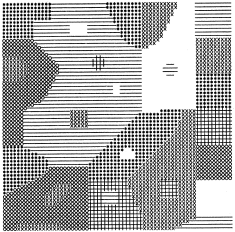


Fig. 3a. True six-colour scene: 64 × 64.

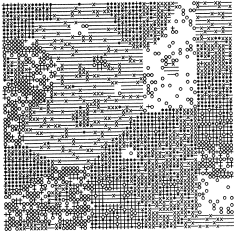


Fig. 3b. Maximum likelihood classifier: 32% error rate

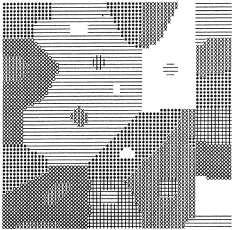


Fig. 3c. ICM reconstruction with $\beta \uparrow 1.5$: 1.2% error rate.

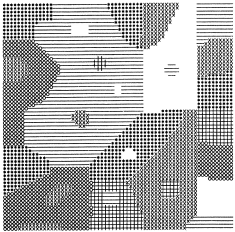


Fig. 3d. ICM reconstruction with β estimated: $\hat{\beta} = 1.80$: 1.1% error rate.

Dirty pictures

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".

Iterated conditional modes

True scene (unobservable) x; observed digital image y.

Recover x by iteratively maximising the posterior local characteristics

$$p(x_i|y, x_{S\setminus i})$$

... a 'Gauss-Seidel' approach. Empirically exhibits superior performance relative to (expensive) MAP estimator

$$\text{argmax}_x p(x|y)$$

MCMC: key papers

Spatial statistics and Bayesian computation (with Discussion). *Journal of the Royal Statistical Society B* (1993) (with Peter Green).

Bayesian computation and stochastic systems (with Discussion). *Statistical Science* (1995) (with Peter Green, David Higdon and Kerrie Mengersen).

Contributions to MCMC

- partial conditioning (extends multigrid Swendsen-Wang)
- randomised proposals (explains adaptive rejection Metropolis sampling)
- Langevin-Hastings (MALA)
- sequential MCMC prediction
- simultaneous credible regions
- promotion/adaptation of statistical physics ideas

Monte Carlo testing: key papers

Simple Monte Carlo tests for spatial pattern. *Applied Statistics* (1977) (with Peter Diggle).

Generalized Monte Carlo significance tests. *Biometrika* (1989) (with Peter Clifford).

Sequential Monte Carlo p-values. *Biometrika* (1991) (with Peter Clifford).

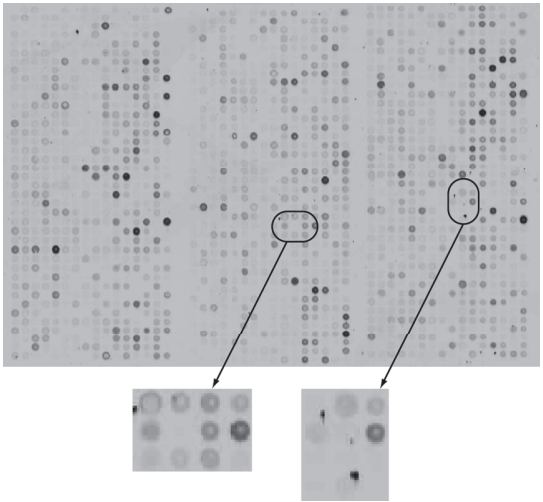
Applications

- medical imaging
- remote sensing
- microarrays
- agricultural field trials
- geographical epidemiology
- biostatistics
- social networks
- ecology, etc.

Microarrays: key paper

Probabilistic segmentation and intensity estimation for microarray images. *Biostatistics* (2006) (with Raphael Gottardo, Matthew Stephens and Alejandro Murua).

Microarrays



Raw image from cDNA array from an HIV experiment: the paper devises and investigates a probabilistic approach to simultaneous segmentation and intensity estimation, implemented using EM/ICM algorithms

Agricultural field trials: key papers

Statistical analysis of field experiments using neighbouring plots. *Biometrics* (1986) (with Rob Kempton).

Bayesian analysis of agricultural field experiments (with Discussion). *Journal of the Royal Statistical Society B* (1999) (with David Higdon).

Field trials: key ideas

- linear models for yield, grounded and informed by well-understood practical context
- adjustment for variation in fertility using neighbouring plots
- decomposition $Y = F\psi + T\tau + \text{error}$
- Bayesian formulation, MCMC computation
  - simplifies interpretation (e.g. ranking, selection)
  - allows complex formulations
  - hierarchical t-formulation (outliers, jumps in fertility)

Field trials

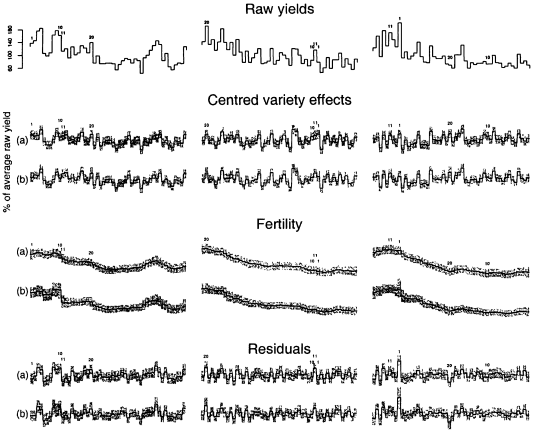


Fig. 1. El Batán variety trial: yields and additive decompositions into variety, fertility and residual effects under (a) Gaussian and (b) hierarchical-t formulations, with the same scale throughout: shaded regions provide pointwise 90% credible intervals; locations of varieties 1, 10, 11 and 20 in each replicate are identified

Geographical epidemiology: key papers

The detection of clusters in rare diseases. *Journal of the Royal Statistical Society A* (1991) (with James Newell).

Bayesian image restoration, with two applications in spatial statistics (with Discussion). *Annals of the Institute of Statistical Mathematics* (1991) (with Jeremy York and Annie Mollié).

Modelling risk from a disease in time and space. *Statistics in Medicine* (1998) (with Leo Knorr-Held).

Space-time analysis of lung cancer incidence

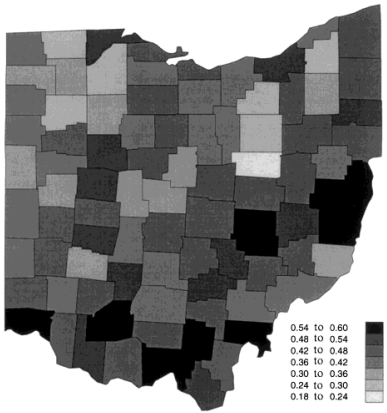


Figure 1. Crude annual death rate $\times$ 1000 for each county in Ohio

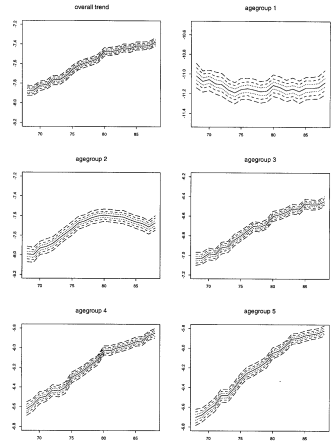
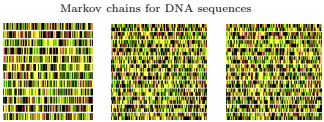


Figure 5. Medians and 50, 80 and 95 per cent credible intervals for the overall time trend and for the aggregate of this with each of the five age group effects

Disease counts treated as independent Binomial, with logit probabilities modelled as `year+age*time+ gender*race*time+ covariate+county`, with `county` a Gaussian intrinsic autoregression plus noise.

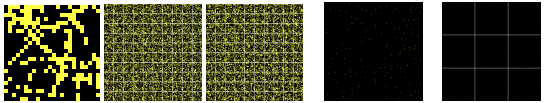
Last seminars



Constrained Monte Carlo and Some Applications

JULIAN BESAG

Department of Statistics, University of Washington, Seattle, USA (emeritus)
Department of Mathematics, University of Bristol, UK (honorary)



1

Last seminars

Agenda

A statistician plays Sudoku

... and defends p–values in exploratory data analysis.

Simple Monte Carlo p–values.
Examples

Markov chain Monte Carlo (MCMC) p–values.
Applications

MCMC for p–values in multi–dimensional contingency tables
Applications

Mobility and irreducibility in constrained sample spaces.
Applications

Social networks : Markov random graphs : MCMC p–values.
Application

2

Last seminars

Continuum limits of Gaussian Markov random fields : resolving the conflict with geostatistics

JULIAN BESAG

Department of Mathematical Sciences, University of Bath, England
emeritus *Department of Statistics, University of Washington, Seattle, USA*

Joint work with DEBASHIS MONDAL

Department of Statistics, University of Chicago, USA
formerly *Department of Statistics, University of Washington*

Oxford, 23 October, 2008

1

Last seminars

Agenda

- **Hidden Markov random fields (MRFs)** and **some applications**.
- **Geostatistical** versus **MRF** approach to **spatial data**.
- Describe simplest **Gaussian intrinsic autoregression** on 2–d **rectangular array** and its **exact** and **asymptotic variograms**.
- Describe **de Wijs process** and its **exact** and **approximate variograms**.
- Reconcile **geostatistics** and **Gaussian MRFs** via **regional averages**.
- **Generalizations** and **wrap–up**.

2

Last seminars

Wrap up

- **Gaussian Markov random fields** are alive and well !!
- **Precision matrix** of **Gaussian** MRFs **sparse** $\Rightarrow$ **efficient** computation.
- **Regional averages** of Gaussian MRFs $\stackrel{\text{rapid}}{\longrightarrow}$ continuum **de Wijs** process.
- **Reconciliation** between Gaussian MRF and original geostatistical formulation.
- **Empirical evidence** for **de Wijs** process in **agriculture** :
  P. McCullagh & D. Clifford (2006), "Evidence of conformal invariance for crop yields", *Proc. R. Soc.* A, **462**, 2119–2143.
  Consistently selects **de Wijs** within **Matérn** class of **variograms** (25 crops !).
- **de Wijs** process also alive and well and can be fitted via **Gaussian MRFs**.
- **de Wijs** process has separate life as **Gaussian free field** in statistical physics.

- Webpage: sustain.bris.ac.uk/JulianBesag/tributes
- Email: P.J.Green@bristol.ac.uk

Honours

1983	RSS Guy medal in Silver
1984	Member of the International Statistical Institute
1991	Fellow of the Institute of Mathematical Statistics
2001	Chancellor's medal, University of California
2004	Fellow of the Royal Society

At the helm of his yacht *Annie* in 2006