



Board of the Foundation of the Scandinavian Journal of Statistics

Topics in Spatial Statistics [with Discussion, Comments and Rejoinder]

Author(s): Nils Lid Hjort, Henning Omre, Marianne Frisén, Fred Godtliebsen, Jon Helgeland, Jesper Møller, Eva B. Vedel Jensen, Mats Rudemo, Henrik Stryhn

Reviewed work(s):

Source: *Scandinavian Journal of Statistics*, Vol. 21, No. 4 (Dec., 1994), pp. 289-357

Published by: [Blackwell Publishing](#) on behalf of [Board of the Foundation of the Scandinavian Journal of Statistics](#)

Stable URL: <http://www.jstor.org/stable/4616322>

Accessed: 02/03/2012 06:41

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at

<http://www.jstor.org/page/info/about/policies/terms.jsp>

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.



Blackwell Publishing and *Board of the Foundation of the Scandinavian Journal of Statistics* are collaborating with JSTOR to digitize, preserve and extend access to *Scandinavian Journal of Statistics*.

<http://www.jstor.org>

Topics in Spatial Statistics*

NILS LID HJORT

University of Oslo

HENNING OMRE

Norwegian Institute of Technology

ABSTRACT. An overview is given over a fair range of topics within spatial and spatial-temporal statistics. The theory presented is motivated by and illustrated with actual applications to real world problems. We describe and discuss models for three basic types of spatial processes: continuous random surfaces, mosaic phenomena, and events-against-background processes. Various combinations of these sometimes occur naturally in applications, like Gaussian noise on top of a Markov random field in image restoration problems. Some of these combinations are also discussed. The applications we discuss are drawn from the areas of medical image analysis, pollution monitoring, characterisation of oil reservoirs, estimation of fish and whale stock, forestry surveillance via satellite, statistical meteorology, and symbol recognition.

Key words: Bayesian methods, covariance function, event processes, hidden Markov fields, image restoration, Kriging, marked point processes, Markov random fields, parameter estimation, pseudo-likelihood, quasi-likelihood, semi-Markov random fields, spatial classification, spatial sampling strategy, stochastic simulation

1. Introduction

1.1. What's special about spatial?

Spatial statistical problems call for evaluation by exploratory data analysis, prediction and classification, simulation, and confirmatory statistics, and are accordingly in these respects well within traditional statistics. To pin-point differences, consider a spatial process $\{Y(x): x \in D\}$ with some m -dimensional $Y(x)$ defined over some s -dimensional spatial or spatial-temporal reference region D . We have encountered $s = 2, 3, 4$ in our applications. Traditional statistical dependence between variables may occur in the Y -space, while the spatial reference x allows for dependencies in the reference dimensions. The latter aspect in particular is challenging from a stochastic modelling point of view and usually adds complexity to the sampling and estimation stages. The presence of a reference variable also points to the importance of scale and transformations between different scales.

Spatial phenomena are complicated to understand and model. The objective of a study is often to evaluate characteristics of a single realisation $\{y(x): x \in D\}$, for example, the hydrocarbon present in one particular petroleum reservoir. In this setting the data points $y(x_1), \dots, y(x_n)$ are “non-repeatable” and the observations may be dependent through their spatial locations. In traditional statistics the underlying assumptions usually include repeatability in the form of (nearly) independent and (nearly) identically distributed observations. In spatial statistics some sort of “pseudo-repeatability” is often obtained by postulating various forms of spatial stationarity. Typical model assumptions could be a constant mean $EY(x) = m$ and/or a stationary covariance function $\text{cov}\{Y(x_1), Y(x_2)\} = C(x_1 - x_2)$. There are parallels to modelling of time series, but spatial problems are often a

*Based in part on invited forum lectures at the Nordic Conference on Mathematical Statistics, Odense, June 1990.

full degree more complicated, due to lack of ordering in the reference space, the fact that border areas of the reference domain often constitute a large proportion of D , and the large variety of sampling options which appear when the dimensionality of D increases. The ergodicity assumption needed in both time series and spatial statistics is much harder to justify in the latter case, since the dependence structure often is strong and D seldomly, with reasonable imagination, can be extended to infinity. Spatial correlation also tends to die out much more slowly than for a dependent one-dimensional process such as time series.

In easier i.i.d.-type problems there is always a Glivenko–Cantelli type theorem which says that one can be ambitious and fit even sophisticated realistic models; the problems are “one-dimensionally” tied to a couple of underlying distributions and nothing else, and the information content of a data set is sufficient to assess intricate features of these. In complex spatial problems, on the other hand, the information content is more thinly spread out in a much larger tapestry of inter-woven problems. Just think of inferring the underlying probabilistic structure of a two-dimensional continuous random function from just observing its values in a finite number of locations for a single realisation!

We have mentioned that sampling of spatial phenomena has several special features. The sampling support B is defined as the domain of the volume over which the sampling is averaged, i.e. $y_B(x_i) = |B|^{-1} \int_B y(x_i - u) du$, writing $|B|$ for the volume (or area) of B . A sampling support of zero is in many cases impossible. In petroleum applications, observations in wells are made on about $0.03 \times 0.03 \times 0.05 \text{ m}^3$, which is of approximate zero support relative to the extent of the reservoir which is typically of size $2500 \times 5000 \times 100 \text{ m}^3$. Seismically-collected data have a considerable support, however, approximately $100 \times 100 \times 10 \text{ m}^3$, and this must be accounted for when the two types of data are combined.

Data representativity in the form of a random sampling hypothesis is usually a basic assumption in traditional statistics. Sampling in non-regular regimes is frequent in applications of spatial statistics, however. “Preferential sampling” of some sort frequently occurs in practice, planned or un-planned. One is of course tempted to use available data and knowledge about spatial continuity in order to confirm favourable areas with extreme values of the process. This makes sense from most points of view except from that of traditional statistics. In the petroleum industry one is using seismic data in order to locate the first well in a prospect such that the chances for hitting oil is maximised. The fact that each exploratory well costs about US\$ 30 million makes preferential sampling almost mandatory. Consequently one particular challenge of spatial statistics is to correct for preferential sampling when, for example, predicting the total hydrocarbon volume.

The spatial dependence entails that the information content in each sample may vary. In image analysis and remote sensing, where regular sampling occurs, the redundancy in the sometimes enormous data set is usually substantial. The amount of information in the set of observations is often much smaller than what traditionally would have been anticipated.

Measurements of the variable of interest are often expensive or complicated to obtain. Spatial problems frequently involve indirect measurements, like seismic data, in order to improve the spatial coverage of the data. This calls for multivariate spatial models, and often for multivariate calibration methods.

The spatial reference in $\{Y(x): x \in D\}$ provides good opportunities for stochastic modelling. The fact that sampling is scarce and that the spatial dependence complicates the exploratory data analysis renders model verification a very difficult task. When preparing a development plan for a North Sea oil reservoir only about $10^{-8} \%$ of the reservoir volume is directly observable from well data. Geological experience and indirect seismic data are other sources of knowledge, and make the evaluation possible. The verification of the model must often be based on experience.

In typical spatial models with extensive dependence between observations one can only seldomly find explicit closed-form parameter estimators with good efficiency properties. Usually one has to rely on model fitting to the observations through iterative procedures and cross-validation. Note that jackknife and bootstrap procedures are difficult to apply in spatial problems because of the complex dependence structure.

We have so far primarily discussed spatial problems. Problems containing both spatial and temporal elements are considered with increasing interest, however. In problems of this type the samples are often abundant, particularly in the time dimension. This is caused by the increasing use of automatic sensors, and again there are often difficulties stemming from a high degree of data redundancy. The interdependence structure is even more complicated than for spatial problems, and even simple exploratory data analysis may turn out to be very complicated.

1.2. Three ways of applying statistics

Let us return to “general statistics” and its philosophy and use in spatial statistics problems. It seems appropriate and convenient to distinguish between three ways of applying statistical methodology.

- (i) *Exploratory statistics* is primarily concerned with the observations $y(x_1), \dots, y(x_n)$ and their characteristics. When exploring them a minimum of model assumptions is enforced. Usually only some statements about representativity are made, so that data displays and meaningful plots associated with summary statistics can be provided. Exploratory statistics is an important but still underused part of applied statistics. Its uses include generating hypotheses justifying model choices.
- (ii) *Predictive statistics* is primarily concerned with prediction and classification of realisations, either because the actual realisation is not observed or because it is observed with noise. An example of the former is spatial interpolation from $y(x_1), \dots, y(x_n)$ to prediction of $y(x)$ in other positions, perhaps with an uncertainty measure included, and an example of the latter is image restoration. Model assumptions are made only to meet the objective of prediction. In other words, the stochastic model is often constructed more for reasons of pragmatism and convenience than for ambitiously and realistically describing the phenomenon under study. Thus a model could have only vague, intuitive connections to the physical phenomenon. Some clear relations from phenomenon to model are of course preferable, since this eases the justification of the model towards the users, but in the end the quality of the model is judged solely by its predictive success. This viewpoint, whereby statisticians fine-tune parameters of algorithms rather than estimate parameters of models, can of course be adopted also in more traditional frameworks of perhaps one-dimensional independent or nearly independent realisations of some phenomenon, but is, we feel, particularly relevant for complex spatial problems. Model fitting is frequently made by cross-validation based on the available observations. Prediction and classification methods find numerous uses in all sectors of applied statistics.
- (iii) *Confirmatory statistics* is mostly occupied with properties of the underlying phenomenon $\{Y(x): x \in D\}$. It requires a stochastic model which in a stronger sense than with predictive statistics truly reflects the key features of the phenomenon, and model parameters must be interpreted in terms of it. The observations $y(x_i)$ are used to express the significance of these parameters. An example is the test for significant correlation between permeability and porosity in sandstone petroleum reservoirs. The quality of this statement is crucially dependent on the validity of the model. Confirmatory statistics are used primarily in such scientific contexts.

The problems addressed in this article will utilise spatial statistics in a predictive setting. Stochastic modelling constitutes a considerable part of the studies presented. This is necessary in order to integrate the different types of available information for the purpose of prediction or classification. Expert experience often constitutes an important source of information, which sometimes invites Bayesian and empirical Bayesian approaches. The Bayesian formalism has proved useful in pragmatic modelling, and since formal testing is seldom performed the disadvantages of the formalism are seldom exposed. Large data volumes with complicated intercorrelation structures and computer intensive solution algorithms are other characteristics.

1.3. Stochastic modelling

Phenomena that vary in space and/or time are frequently observed in nature. The use of stochastic models and statistics in surveying such phenomena has proven useful. Stochastic modelling constitutes the artistic part of the analysis, and some rules of thumb should be kept in mind. The model formulation must be tailored to the question to be answered. Classification of discrete objects, predictions of a continuous surface, and identification of discontinuities require different model formulations. If extensive knowledge about the phenomenon to be evaluated is available it should be used in the modelling. The scale at which the model is valid has to be specified. Consider a porous medium like sandstone; at micro scale a discrete spatial model based on pores and sand grains would be suitable, while on macro scale porosity could be represented by a continuous spatial model. The amount of available data will also influence the modelling. The problem of overfitting is well recognised in statistics, and in spatial-temporal settings the number of parameters is often large and one encounters redundancies in the data. This makes the problem even more crucial. Bayesian approaches with prior qualified guesses on parameter values based on phenomenological information will reduce the problems of overfitting. As previously mentioned, data analysis is often complex and efficient model estimators of known form are rarely available in spatial-temporal settings. Hence the possibilities for model verification and for interpretation of and estimation of parameters should be taken into consideration when choosing the statistical model.

Spatial problems appear with complex interdependence structures, and a large variety of spatial models can be imagined. In this presentation a broad division into three natural main model classes has been made, and this corresponds roughly to the three types of models most frequently encountered in spatial statistics literature. The division is into classes of well-defined mathematical objects, but is mainly motivated by the form of real spatial problems and by the form of the available data sources.

- (i) *Models for continuous random surfaces*, say $\{Z(x): x \in D\}$ with $Z(x)$ in some m -dimensional Euclidian space. The model most frequently used is the Gaussian random function model, perhaps after an initial scale-transformation like taking logarithms. It maintains most of the favourable properties from non-spatial models when introducing higher dimensional spatial references. Typical tasks to be carried out include interpolation by estimated conditional expected value $E\{Z(x) | z(x_1), \dots, z(x_n)\}$ and data-conditional simulation of $\{Z_{\text{sim}}(x): x \in D | z(x_1), \dots, z(x_n)\}$.
- (ii) *Models for mosaic phenomena*, say $\{L(x): x \in D\}$ with $L(x) \in \{1, \dots, K\}$. The Markov random fields constitute the most popular class of models, but tessellation techniques are also used. These form an extremely large class of models and only small parts of their potential have been explored. Unfortunately, the nice mathematical properties found in the one-dimensional case are not maintained when higher-

dimensional spatial references are introduced. The lack of ordering causes this. Most applications include stochastic simulations, as a means in itself or as some intermediate block, with the Metropolis algorithm or Gibbs sampler providing ways of generating realisations.

- (iii) *Models for events-against-background processes*, say $\{(x_1, S_1), \dots, (x_n, S_n)\}$, in which S_i is a set of attributes assigned to reference location x_i . The models most frequently used are related to the theory of marked point processes. An example could be a simultaneous model for locations and heights of trees in a forest. This model can also be defined in a general manner, but usually only pairwise dependencies of marked points are modelled. Few exact analytical results are available for processes outside the simpler Poisson type ones. Statistical analysis of this type of model is typically carried out through simulation using variations of Ripley–Kelly's spatial birth-and-death algorithm.

1.4. The present article

The material is organised as follows. Section 2 presents a generous list of application examples, sorted into problem areas. The emphasis is on describing problems and modelling ideas, and is not on “solutions”. The applications have been chosen from a much longer list of projects we have worked on, with fellow statisticians at the Norwegian Computing Centre and surroundings and with clients. We have strived to represent the most important dimensions in this high-dimensional space of all spatial statistics applications. We have partly been guided by the degree of problem-solving success as criterion but also by the methodologically inclined statistician's view of what constitute interesting models and interesting problems. One of our aims is to show to the statistical community the kind of statistical problems that are currently deemed important to user groups. The application examples presented come from projects that are actually paid for by clients. Judging the usefulness of applied statistical models, methods and expertise by the willingness of users to spend money on them is not an uninteresting yardstick.

Section 3 presents basic methodology, introduces the most useful stochastic models, and discusses ways of analysing them. In particular models for the three main types of phenomena noted above are discussed, as well as a couple of “combinations” where two different models work together. Four of section 2's list of applications are returned to in section 4 for a more complete and more careful treatment, and suggested solutions to the actual real problems are described. Finally section 5 gives some concluding remarks and points to some topics for future work.

This is partly a review of several topics in spatial statistics with a broad range of examples. Some basic references about models in and uses of spatial statistics include Matérn (1960), Yaglom (1962), Matheron (1965), Journel & Huijbregts (1978), Diggle (1983), Ripley (1981, 1988), Stoyan *et al.* (1987), Cressie (1991), and Walden & Guttorp (1992). For the convenience of some readers we point out here what are supposed to be “new contributions”, or perhaps only modest new insights into the use of old methods, in our article: the extended “conjugate family” analysis for Bayesian Kriging in section 3.1.3; some new reliable simulation methods for Gaussian surfaces in section 3.1.4; the quasi-likelihood method in section 3.1.5 for estimating covariance function parameters; comparison of maximum likelihood and maximum pseudo-likelihood for Markov chains in section 3.2.2; the semi-Markov type random field models of section 3.2.5; the ways of imposing global constraints on realisations from Markov random fields and marked point processes in sections 3.2 and 3.4 respectively; and generalisations of Geman and Geman

and of Besag methods in image restoration problems with correlated noise, in section 3.4. In addition we hope that the ways by which we approach and solve some of the real world problems in sections 2 and 4 contain some novel ideas in the respective problem areas.

2. Range of applications

In the following a collection of application examples are briefly described. Four of these are returned to in section 4 for a more complete treatment, leading to suggested solutions to the actual real problems.

2.1. Medical image analysis

2.1.1. Tumor identification (Lundervold *et al.*, 1988)

Identification and classification of tumors in the human brain is obviously a problem of great importance. Magnetic resonance equipment provides the possibility for indirect measurement of various characteristics of the brain with three-dimensional spatial references. The data can be collected without surgery, hence minimising the chance of complications. The observations are indirect and the signal to noise ratio is low. A model based on hidden Markov random field theory with Gaussian noise is often used for segmenting the three-dimensional brain into various pathological units. The pathological units are modelled by requiring $p(X_{ij} = k | \text{rest of image})$ to depend upon units in the 5×5 neighbourhood of pixel (i, j) only. Here X_{ij} denotes pathological unit type at pixel (i, j) . The noise component is usually of the white noise type, or it may be spatially auto-correlated within some neighbourhood. See sections 3.2 and 3.4. Usually no direct observations will be available, which means that unsupervised classification of the units must be made. Spatial models are required because of the areal extent of the pathological units and the spatially correlated noise.

Experiences with medical image analysis have so far been encouraging, and the methods will hopefully be in commercial use in the near future. From a statistical point of view the traditional Markov random field theory has several shortcomings as a model for the pathological units. The problem surfaces in changes of scale and in parameter estimation, where good estimators for model parameters are hard to construct. Further studies of other models that are more directly suitable for segmentation are needed.

2.1.2. Identification of heart dysfunction (Taxt *et al.*, 1990; Storvik & Switzer, 1992)

Heart defects may appear as reduction in pumping capacity, reduced volume pumped, and as decline in elasticity of the heart walls. The dynamic behaviour of the heart can be observed by repeated three-dimensional grey-tone ultra-sound images at 25–35 Hz. The signal-to-noise ratio in each image is normally very poor, hence the time repetitions must be utilised. See section 4.2 for further discussion.

2.1.3. Noise reduction in nuclear magnetic resonance imaging (Godtliebsen, 1989)

The common technique for reducing noise in NMR images is to take several measurements on the same slice and then average. This is time-consuming and expensive, and the patients sometimes move during acquisition time, thereby introducing additional noise. Hence a natural challenge is to devise statistical noise reduction algorithms that work on a single slice. One approach is to model the observed image as $y_i = x_i + \varepsilon_i$ in pixel i , where the collection

of x_i s come from some Markov random field and the ε_i s are independent Gaussian zero mean noise. Studies indicate that the latter assumption is quite acceptable. The Markov random field assumption is less realistic, but can be used to derive image enhancement and image noise reduction techniques. Examples of such methods are in section 3.4. Another method which can be motivated by the model assumptions is to replace observed grey level in a pixel with some weighted average of grey levels over the pixel and neighbouring pixels, with weights determined by the spread of the local data. This can be done in a suitable empirical Bayesian fashion. The study found that sound statistical techniques were able to reduce noise in a single picture with a factor of about three. This particular study was unusual in that a good approximation to the "true scene" was available, taken to be the average of eight consecutive images of the same slice. Accordingly measures of restoration performance could be proposed and compared for different image analysis methods.

2.2. Pollution monitoring

2.2.1. Status of forest (Strand, 1989)

The decline in the quality of forest may be linked to increased pollution, and the Norwegian authorities have initiated an extensive sampling programme. Sampling takes place in more than 2000 sites in a regular $9 \times 9 \text{ km}^2$ grid over Norway every second year. Each site is 100 m^2 and each tree is located and characteristics such as age, size and top density are sampled. The general environment at each site is also carefully sampled. A spatial regression model is used to analyse the data, see section 3.1. The model is $S(x) = \sum_{i=1}^p \beta_i f_i(x) + \varepsilon(x)$, where $S(x)$ is some variable representing status of forest and the $f_i(x)$ s are known explanatory regressor functions such as elevation, soil quality etc. The residuals $\varepsilon(x)$ are treated as a random function and its regional properties are evaluated in order to find differences due to unexplained factors, which could include air pollution. The data analysis exposes reasonably large regional differences, which at this stage are believed to be linked to pollution level. An extended analysis is planned.

2.2.2. Air quality (Halvorsen & Strand, 1987; Høst *et al.*, 1994)

The decline in air quality over Europe is a problem of concern. There are several sources for the pollution, which is transported over long distances. Its solution calls for international cooperation. The European Meteorological Environment Programme has established more than one hundred monitoring stations all over Europe. The air quality is characterised by the volume content of several chemical components, particle density, etc. Sampling is made on a daily basis. Consequently the pollution monitoring can be considered as a time- and space problem, and one approach for evaluation of the problem is further discussed in section 4.4.

2.2.3. Combining satellite data with field data in pollution monitoring (Høst *et al.*, 1989)

Some water quality variables from the Hvaler area in Norway have been analysed on the basis of both hard-to-get direct measurements and easy-to-get satellite data. The satellite data are abundant but of course very "indirect", having a small positive correlation with the water quality variables. The challenge is to build a model that makes it possible to integrate these very different data types. On the basis of a validated model a map of the estimated water quality was produced, along with a map of the estimated uncertainty of the estimate.

2.3. Reservoir characterisation

2.3.1. Seismic depth conversion (Omre *et al.*, 1989; Abrahamsen *et al.*, 1991; Abrahamsen, 1993)

Petroleum is usually trapped under a geologic horizon having non-permeable characteristics, for example shale. Fortunately, these types of horizons can also be identified from seismic data. The seismic data have good spatial coverage and consist of two-way reflection times down to the horizon. Note that the unit here is time, while the geologists are interested in depth to the horizon. By using depth observations in a small number of wells, seismic reflection time data and basic laws of physics, a model for depth conversion can be constructed. This is further discussed in section 4.1.

2.3.2. Simulation of facies architecture (Clemetsen *et al.*, 1989; Hjort *et al.*, 1989; Høiberg *et al.*, 1990, 1992; Omre, 1992; Georgsen & Omre, 1993; Tjelmeland & Holden, 1993)

The petroleum reservoirs in the North Sea appear as heterogeneous in the sense that several units of good and poor quality are packed. The units usually correspond to different rock types of facies. Their packing is according to certain geologic processes. The heterogeneity in the reservoirs has proven to have large impact on the production potential.

The facies architecture is a consequence of the geologic processes, the dynamics of which are partly understood by the geologists. This constitutes the primary base for the modelling. The facies distribution can be observed in the wells, and this provides constraints on the model. Both Markov random fields and marked point processes have been used in modelling the facies architecture, see sections 3.2 and 3.3. The former model postulates that $p(X_{ij} = k \mid \text{all other facies})$ is only dependent upon the neighbourhood facies, with X_{ij} facies type in pixel location (i, j) and $k \in \{1, \dots, K\}$ facies type. Sometimes there are as many as $K = 12$ facies types on the scene. Simulations of pseudo-reality constrained by some known values of x_{kl} are required. Simulation procedures from recent literature seem to converge very slowly, and exploring their properties is a difficult and time-consuming task. The marked point process traditionally applied has density of the form $f_n(m_1, \dots, m_n) = \text{const exp}\{\sum_{i=1}^n b(m_i) - \sum_{i < j} c(m_i, m_j)\}$, with m_i s being the marked points with information on location, size, shape, and facies characteristics. Often global constraints are necessary, and simulated realisations are pushed in the wished-for direction by certain tricks. New model formations of the "semi-Markov" process type have also been studied and seem promising. In order to evaluate the impact of the heterogeneity on the production of petroleum, simulations of fluid flow are performed on realisations of the facies architecture.

2.3.3. Simulation of fractures and faults (Omre *et al.*, 1992a, b)

The petroleum reservoirs in the North Sea are of sedimentary origin, and they have been changed by considerable tectonic activity. This has forced a complicated fracture and fault pattern on to the reservoirs. The location of large fault zones, i.e. those with offset above 10 m, can be observed on the seismic data. The actual break pattern in the zone has been studied by geologists and is found to consist of swarms of smaller faults. This heterogeneity is important for fluid flow. A more thorough presentation of the problem and model appears in section 4.3.

2.3.4. Interpretation of well log data (Bølviken & Helgeland, 1989; Bølviken *et al.*, 1991; Bølviken, 1993)

The petroleum reservoirs in the North Sea are located at a depth of approximately 3000 m. Wells are drilled to penetrate the reservoir in order to collect information about its characteristics. Few direct observations are available even in the wells. Logging tools are lowered down the well, however, and indirect measurements of radioactivity, acoustic reflection, conductivity etc. are collected every 0.25 m in the reservoir zone. From these data the geologists would like to infer the geologic environment from which the reservoir originates. This entails determining the geologic sequences or sequence of rock types down the wells. The problem can be considered as a spatial segmentation problem based on multivariate data from the log tools.

The model is based on hidden and in fact even on hidden hidden Markov random processes with Gaussian noise, see sections 3.1 and 3.4. There are three ordered stages in the evaluation, say $S \rightarrow L \rightarrow X$, with sedimentary processes S creating a sequence of lithofacies L which again influence the responses of the logs X . The cornerstone in the model used is that the unobservable part (S, L) has been generated by a Markov process while L can be observed with white Gaussian noise through X . The geological processes make some sequences more probable *a priori* than others, therefore adding a Bayesian dimension to the model. A general model has been constructed and is expected to be widely applicable, but the set of parameters must be estimated for each reservoir.

2.4. Sea resources

2.4.1. Stock of capelin (Hjort & Murray, 1912; Hjort, 1914; Omre & Sølna, 1991a, b)

The fisheries in Norway are important for employment in the western and northern parts of the country and for the national export volume. The fish resources are renewable, but the reproduction cycle varies among the species. For capelin it is approximately four years in the Barent Sea. While capelin is a consumer of low level organisms such as plankton, it is predated by cod in the winter season. The migration of capelin south to the Norwegian coast for spawning in winter causes the contact with cod. Both capelin and cod are of commercial value, hence a multi-species catching strategy is required.

There are surprising amounts of data available. Acoustic data, indirectly observing the echo-sounding reflected by fish with a three-dimensional reference, are abundant. There are trawl samples as well. For several thousand capelin, multi-dimensional observations of age, length, weight, stadium, etc. with space-time references are available each year. For large amounts of cod, stomach content has been analysed with respect to fraction and volume of capelin. Presently only single-species models are operable, and time is the only reference variable. Multi-species models between capelin and cod are being developed, and both time and space references are discussed. It is at present problematic to verify a significant interaction between the two species based on available observations. A reliable space-time model for the species would compensate for some of the time and space variability, and may contribute to a more reliable analysis. Work is in progress on such matters.

2.4.2. Stock of Minke whale (Hjort & Murray, 1912; Schweder *et al.*, 1990; Schweder & Høst, 1991)

Minke whales are considered an endangered species and have been protected by the International Whaling Commission since 1986. The Norwegian authorities have during the last few years performed surveys in order to estimate the size of the stock. "Official

estimates" have been surprisingly low compared to historical catch successes. A serious downward bias is expected in the predictions since these have been based on the assumption of complete sampling in the surveyed areas. Due to the fact that each whale surfaces only about 30 times per hour and that they can be difficult to detect in rough sea, the actual detection success rate for whales passing close to the survey vessel, $g(0)$, is probably significantly smaller than 1.

There are two stochastic elements in the final estimator of the population size. One is the hazard probability $Q(x, r)$, the probability of sighting a whale surfacing at polar coordinates (x, r) relative to the vessel, given that the whale has not been observed before. The parametric form of $Q(x, r)$ is a subject of continued discussion. Its parameters can be estimated from data provided by test surveys where two vessels were run in parallel and covered the same area. The second random element is the surfacing frequency for whales. This has been reproduced by simulation from a spatial Poisson process. From this model the sampling success rate $g(0)$ was found to be approximately 0.5, which means that about one out of two whales were observed. Thus the predicted number of whales is about twice as large as first anticipated.

Methods developed here are partly of a general nature, and aim at being able to integrate very different types of data (viz. "micro" and "macro" data) in a consistent and meaningful framework. They should find applications in other areas as well.

2.5. Other areas of application

2.5.1. Mapping of seabed: spatial sampling strategy to find all shallow areas (Helgeland et al., 1984)

Let $Z(x)$ be depth to the seabed in geographical location x . Mapping of $Z(x)$ based on point sampling is another problem of spatial interpolation, see section 3.1. Suppose however that it is considered important to find all shallow banks, say where $Z(x)$ is smaller than some level u . Term sets of the type $\{x: Z(x) < u\}$ by Z_u -areas. A question of spatial sampling strategy is therefore: what is a good regime for detecting all Z_u -areas, and what is the probability of not detecting such an area? A solution based on approximate shapes of excursion sets for Gaussian random fields is given in the above reference.

2.5.2. Automatic recognition of handwritten symbols (Hjort, 1986; Hjort & Taxt, 1988; Pripp, 1990)

Automatic recognition of printed or handwritten symbols is an established field with several well-explored approaches. Among these are several which are statistical in nature. The usual method comprises two main steps. The first is to extract a feature vector for the symbols, typically of dimension 10 or less. The second is to model the behaviour of feature vectors for each symbol type, and then estimate parameters, perhaps in cheap semi-automatic ways. Finally statistical discriminant analysis is used on future symbols.

A more direct approach would be to model the statistical behaviour of the symbols themselves. We have some experience with modelling the boundaries of symbols as random closed curves in the plane, giving rather successful rates of correct classification. This is perhaps not a genuine spatial example since we merely model one-dimensional objects. But the following approach is spatial. Suppose the candidate symbol is digitised to form 0s and 1s on a rectangular grid. Thus a hand-drawn "8" could be digitised on a 20×20 grid and be represented by the resulting collection of 4000 0s and 1s. Then a possibility is to model the mosaic process of 1s on this lattice as a Markov random field, see section 3.2. Each

symbol class (say a hand-written “8”) has its own mrf specification of the conditional probability of having a “1” in pixel location (i, j) , given the rest of the image. It is of a certain form involving various “award functions” designated by the modeller to encourage or disencourage certain types of local behaviour, and various parameters, some of which may be class-dependent and may vary over the scene. It is sometimes fruitful to impose global constraints too. The mrf parameters can be estimated from data by maximising the product of individual pseudo-likelihoods, see section 3.2. Finally the estimated models are used to construct a classifier.

2.5.3. Meteorology: combining new satellite data with other information sources to improve prognoses (Homleid, 1992)

The Norwegian Institute of Meteorology uses two numerical weather prognosis models for the atmosphere. Input data for such models are observations from ships, radio buoys and sondes, and land-based stations. One is also interested in exploiting satellite data, for example temperature and humidity profiles processed from the TOVS satellite, to build better models for prognosis. Several methods have been tried out but so far the results are not convincingly better. Better ways of combining these very different data sources are currently being explored. The task seems to require (i) simply assessing the current data quality from the satellite, (ii) constructing a successful spatial-temporal statistical model, drawing on both meteorology physics and empirical statistics, (iii) estimating necessary parameters and implementing prognosis formulas, and (iv) evaluating the performance compared to existing methods. Although satellite data obviously add important information to the problem, the improvements in prognosis quality by their inclusion seems to be rather small with the existing prognosis techniques. Research that aims at refining the statistical model formulations, and at a better statistical understanding of why the current improvements are so small, is under way.

3. Theoretical tools

This section presents basic statistical theory that has been developed to solve problems like those listed in section 2. As mentioned above a fruitful division of the various stochastic processes encountered is into continuous random surfaces, finitely-valued or mosaic phenomena, and events-against-background processes. Subsections 1, 2 and 3 study these three basic types. In many problems different data sources may have to be combined, and some combination or other of the three basic model types is called for. Some situations of this sort are discussed in section 3.4.

3.1. Continuous random surfaces

The simplest spatial statistical model capable of describing interesting continuous or near continuous random surfaces is one with some smooth trend surface plus a spatially correlated Gaussian residual process. This model is introduced in section 3.1.1 below. Various aspects of such models are discussed, including theory for spatial interpolation, Bayesian Kriging, simulation, and for estimating parameters in spatial covariance functions.

3.1.1. The basic model

Let $z = z(x)$ be a continuous or nearly continuous surface defined over some domain D of x -values, for example a rectangle in the plane. Suppose $z(x_i)$ -data on $z(\cdot)$ are collected in n

distinct locations x_1, \dots, x_n , and that some problem of interest can be phrased in terms of $z(\cdot)$, like that of spatial interpolation. The spatial statistical way of approaching such problems is to view $z(\cdot)$ as a realisation of a stochastic process $Z(\cdot)$. The idea is to translate prior knowledge to a suitable class of models for $Z(\cdot)$, typically viewed as a smooth trend surface plus spatially correlated residual, use data to estimate parameters, and answer the original $z(\cdot)$ -question under the model assumption and given all available information. To be specific, suppose that $Z(\cdot)$ is Gaussian with regression type trend surface plus zero mean Gaussian residual, say

$$Z(x) = m(x, \beta) + \varepsilon(x) = \sum_{j=1}^p \beta_j f_j(x) + \varepsilon(x). \quad (3.1)$$

Here the $f_j(x)$ s are known regressor functions ($f_1(x)$ would typically be the constant 1), the β_j s are coefficient parameters, and the covariance function

$$\text{cov}\{Z(x), Z(y)\} = \text{cov}\{\varepsilon(x), \varepsilon(y)\} = \sigma^2 K(x, y) \quad (3.2)$$

describes the variability and the degree of spatial continuity of the residual process. One often postulates shift invariance, so that $K(x, y)$ is of the form $K_0(x - y)$ for appropriate $K_0(\cdot)$ function, and in such cases it is convenient to choose $K(x, x) = K_0(0) = 1$ so that $\text{var } Z(x) = \sigma^2$. The random function is isotropic if in addition $K(x, y)$ only depends on the distance $\|x - y\|$, as in the popular case $K(x, y) = \exp\{-c\|x - y\|\}$.

At this point it is worth noting that the covariance function is only defined when $Z(x)$ has finite variance. A richer class of measures for second order spatial characteristics is the so-called semi-variogram

$$\gamma(x, y) = \frac{1}{2} \text{var}\{Z(x) - Z(y)\}, \quad (3.3)$$

which requires only the variances of differences to be finite. When K is shift invariant and $\text{var } Z(x) = \sigma^2 < \infty$ one has $\gamma(x, y) = \sigma^2\{1 - K(x, y)\}$. The idea of and value of requiring finite variances of only certain linear combinations of $Z(\cdot)$ is developed in the theory of intrinsic random functions, see remarks at the end of the following section. The possible choices for semi-variogram functions are linked to the choice of p and $f_j(x)$ s together with the requirement of producing non-negative prediction variances; see also remarks below. We have chosen to present most of the general Gaussian random function theory in terms of spatial covariance functions.

3.1.2. Spatial interpolation by universal Kriging, and its precision

Suppose interpolation is called for. Let x be a new location point, and let us follow the programme above. Under the Gaussian assumption

$$\begin{pmatrix} Z(x) \\ Z_{\text{dat}} \end{pmatrix} \sim \mathcal{N}_{n+1} \left\{ \begin{pmatrix} f(x)' \beta \\ F\beta \end{pmatrix}, \sigma^2 \begin{pmatrix} K(x, x) & k' \\ k & K \end{pmatrix} \right\},$$

in which K is the $n \times n$ matrix of $K(x_i, x_j)$, k and Z_{dat} are the vectors with components respectively $K(x, x_i)$ and $Z(x_i)$, and finally F is the $n \times p$ matrix whose i th row is $f(x_i)' = (f_1(x_i), \dots, f_p(x_i))$. Hence

$$Z(x) | \text{data} \sim \mathcal{N}\{m(x, \beta) + k'K^{-1}(Z_{\text{dat}} - m(\beta)), \sigma^2(K(x, x) - k'K^{-1}k)\},$$

writing $m(\beta)$ for the vector of $m(x_i, \beta)$. Of course $m(\beta) = F\beta$ in the present case, but the notation is meant to suggest its natural generalisation to other regression functions.

Specification and estimation of $K(\cdot, \cdot)$ is discussed in section 3.1.5 below. Suppose for now that a covariance function has been decided on. The natural estimator of β then emerges by minimising $(Z_{\text{dat}} - F\beta)'K^{-1}(Z_{\text{dat}} - F\beta)$. This is the weighted least squares as well as the maximum likelihood principle, when the covariance function is assumed known. The result is

$$\hat{\beta} = HF'K^{-1}Z_{\text{dat}} \quad \text{where } H = (F'K^{-1}F)^{-1},$$

constituting an unbiased estimator with covariance matrix $\sigma^2 H$. The spatial interpolator used in the end is the estimated mean value of $Z(x)$ given the data, that is

$$\hat{Z}(x) = m(x, \hat{\beta}) + k'K^{-1}(Z_{\text{dat}} - m(\hat{\beta})) = f(x)' \hat{\beta} + k'K^{-1}(Z_{\text{dat}} - F\hat{\beta}). \quad (3.4)$$

Note that it is an unbiased predictor in the sense of having $E\{\hat{Z}(x) - Z(x)\} = 0$. The interpolation variance, or prediction error, can be shown to be

$$\begin{aligned} \sigma_{\text{pe}}(x)^2 &= E\{\hat{Z}(x) - Z(x)\}^2 \\ &= \sigma^2[K(x, x) - k'K^{-1}k + (f(x) - F'K^{-1}k)'H(f(x) - F'K^{-1}k)]. \end{aligned} \quad (3.5)$$

Note that the mean squared error is computed w.r.t. (the random) $Z(x)$ and not its mean value $f(x)'\beta$, since the intention is to guess $Z(x)$ for the surface under study and not its trend surface. In particular $\sigma_{\text{pe}}(x)^2$ is not the same as $\text{var } \hat{Z}(x)$.

The interpolator (3.4) is itself independent of the scale factor σ^2 , which however is needed to assess the uncertainty as in (3.5). To estimate σ^2 , when some covariance structure $K(\cdot)$ has been decided on, note that $Z_{\text{dat}} - F\hat{\beta}$ has variance matrix $\sigma^2(K - FHF')$, from which it follows that $Q(\hat{\beta}) = (Z_{\text{dat}} - F\hat{\beta})'K^{-1}(Z_{\text{dat}} - F\hat{\beta})$ has mean value equal to σ^2 times the trace of $K^{-1}(K - FHF')$, which is $\text{tr}(I_n - HF'K^{-1}F) = \text{tr}(I_n - I_p) = n - p$ by the usual tricks. Thus $Q(\hat{\beta})/(n - p)$ is unbiased, even in the present setting with correlated data. The maximum likelihood solution, trusting normality, is $\hat{\sigma}^2 = Q(\hat{\beta})/n$. We should stress that σ^2 and $K(\cdot)$ are defined "together" and should be estimated together. Modelling and estimating $K(\cdot)$ is the harder task, see section 3.1.5 below; σ can be estimated as just described for given $K(\cdot)$. It is also usual in geostatistics to estimate σ^2 by cross-validation techniques, see Davis (1973) and Solow (1990).

Spatial interpolation of a random function with drift can be considered from a somewhat different perspective as well, that of choosing an optimal linear combination, under different assumptions about spatial smoothness. Consider the interpolator $Z^*(x_0) = \sum_{i=1}^n c_i Z(x_i)$ at point $x = x_0$, where the weights c_i are to be determined. Unbiasedness, in the sense of $E\{Z^*(x_0) - Z(x_0)\} = 0$, is ensured by the constraints $\sum_{i=1}^n c_i f_j(x_i) = f_j(x_0)$ for $j = 1, \dots, p$. A natural avenue is to minimise the interpolation variance $\text{var}\{Z^*(x_0) - Z(x_0)\}$ under these constraints. This may be formulated as minimising $\text{var}\sum_{i=0}^n c_i Z(x_i)$ under $\sum_{i=0}^n c_i f_j(x_i) = 0$, $j = 1, \dots, p$, where $c_0 = -1$, and the task is solved by the Lagrange technique. The result is in fact $\hat{Z}(x)$ of (3.4), and this Lagrangian way of deriving it is the usual one in the geostatistics tradition. It is called the universal Kriging interpolator, see for example Journel & Huijbregts (1978). Note that the Gaussian assumption, which was used to reach (3.4), is unnecessary in this construction.

One sometimes uses (3.4), in that form or computed by constrained minimisation, with $K(x, y)$ functions that are not genuine non-negative definite covariance functions. The minimum requirements on the K functions for guaranteeing non-negative interpolation variances are that

$$\text{var} \sum_{i=0}^n c_i Z(x_i) = \sum_{i=0}^n \sum_{l=0}^n c_i c_l K(x_i - x_l) = \begin{pmatrix} -1 \\ c \end{pmatrix}' \begin{pmatrix} 1 & k' \\ k & K \end{pmatrix} \begin{pmatrix} -1 \\ c \end{pmatrix} \geq 0$$

for all x_0, x_1, \dots, x_n and all c_1, \dots, c_n satisfying the $\sum_{i=0}^n c_i f_j(x_i) = 0$ constraints above. It is assumed here that $K(x, y) = K(x - y)$.

Consider cases for which $f_1(x)$ is the constant 1, i.e. the trend surface contains a constant β_1 , and $\sum_{i=1}^n c_i = 1$ among other constraints. Non-negative interpolation variances are ensured by $c'Kc \geq 0$ for all c vectors obeying certain constraints, where K is the matrix of $K(x_i - x_j)$. In terms of the semi-variogram function (3.3) the criterion becomes $c'\Gamma c \leq 0$ with constraints on c , where Γ is the matrix of $\gamma(x_i - x_j)$.

Assume in particular that the regressor functions $f_j(x)$ are polynomials in the coordinates of x , of order k or less. The random functions having this property are called “generalised intrinsic functions of order k ”, and the associated function $K(\cdot)$ is termed a “generalised covariance function of order k ”. It is clear that this class of generalised covariance functions is larger than the class of simply non-negative definite functions.

By enforcing stronger assumptions on the form of the expected value one can choose among a larger class of generalised covariance functions. This approach to spatial interpolation is widely used in the geostatistics school, see Matheron (1973). It resembles the “integration approach” used in time series analysis, see Box & Jenkins (1976). It can be shown, however, that interpolations based on assumptions of intrinsic random function hypotheses, of any given order k , are equivalent to those obtained from an appropriate universal Kriging with a Gaussian random function model, see Christensen (1990).

Remark 1. A sound implementation is vital here, since inverting large matrices can be slow and unstable if done directly. Rather than using the mathematically and statistically informative (3.4) and (3.5) a good interpolation package would typically use derived formulae from Choleski triangularisation, see for example Ripley (1981, sect. 4.4). \square

Remark 2. One valid variogram function, among others, is $\gamma(h) = \sigma^2 h^{2H}$, where $0 \leq H < 1$. This family of variogram functions is said to have affine similarity properties and is used to model fractal phenomena, see for example Feder (1988). The H constant is called the Hurst exponent. The Brownian motion process in one dimension is of this type, with $H = 1/2$. \square

Remark 3. Note that $\hat{Z}(x_i) = z(x_i)$ and $\sigma_{pe}^2(x_i) = 0$, that is, the interpolator respects the data points. This is as it should be, since the real interest is interpolation of the actual $z(\cdot)$, rather than its underlying trend surface. In applications of classical non-parametric regression the problem is typically the opposite one of estimating the smooth trend, based on unrelated realisations at different locations. \square

Remark 4. In locations x far from all sampled x_i s the interpolator is close to the trend estimate $m(x, \hat{\beta})$. \square

Remark 5. More robust estimates than the least squares $\hat{\beta}$ could be used as well, without seriously affecting the reasoning or the results. Our arguments can similarly be generalised to cover non-linear models for $m(x, \beta)$, like $\exp(\beta'x)$, without much difficulty. \square

Remark 6. The geostatistical Kriging techniques have been extended to cover multivariate cases, see for example the so-called co-Kriging method of Journel & Huijbregts (1978). Another extension is called factorial Kriging, see Sandjivy (1984), consisting in separating a random function into a smoothly varying trend and a correlated residual term. \square

Remark 7. The term Kriging was originally used for linear predictors, i.e. linear in the data points $Z(x_i)$. A couple of non-linear predictors have also adopted the Kriging name,

however. Disjunctive Kriging is based on a Hermitean expansion of the bivariate characteristics of the random function, thereby extending the familiar correlation framework, see Matheron (1976). Indicator Kriging is based on a discretisation of the univariate variable into a set of linear combinations of indicator variables, hence providing estimates of the quantiles of the conditional distribution, see Journel (1983) and Isaaks & Srivastava (1989). \square

Remark 8. Markovian properties of Gaussian random functions in the one-dimensional case are easily defined and well understood. There are unexpected difficulties with the different possible definitions of Markov-ness in higher dimensions, however, and some of these lead to too restricted classes. See Adler (1981, app) for a review. \square

3.1.3. Bayesian Kriging

In some situations there is rather too little data to do interpolation as above with the wished for precision, but there is some prior knowledge about the trend surface. This invites Bayesian and empirical Bayesian considerations. The following treatment partly extends that of Omre (1987) and Omre & Halvorsen (1989). Other relevant references are Pilz (1991) and Le & Zidek (1992).

Let the model be as in (3.1) and (3.2) conditionally on β , and suppose β is given some Gaussian prior distribution, which we parameterise as $\mathcal{N}_p\{\beta_0, \sigma^2 T\}$. For the moment we take σ^2 to be known. The joint distribution of β and data Z_{dat} is easily found to be

$$\begin{pmatrix} \beta \\ Z_{\text{dat}} \end{pmatrix} \sim \mathcal{N}_{p+n} \left\{ \begin{pmatrix} \beta_0 \\ F\beta_0 \end{pmatrix}, \sigma^2 \begin{pmatrix} T & TF' \\ FT & K + FTF' \end{pmatrix} \right\}.$$

In particular the posterior distribution of β is still normal. After using the convenient matrix identity $(K + FTF')^{-1} = K^{-1} - K^{-1}FG_T F' K^{-1}$, where $G_T = (H^{-1} + T^{-1})^{-1}$, one finds

$$\begin{aligned} \tilde{\beta} &= E\{\beta \mid \text{data}\} = \beta_0 + TF'(K + FTF')^{-1}(Z_{\text{dat}} - F\beta_0) = G_T H^{-1} \hat{\beta} + (I - G_T H^{-1})\beta_0, \\ \text{var}\{\beta \mid \text{data}\} &= \sigma^2 \{T - TF'(K + FTF')^{-1}FT\} = \sigma^2(I - G_T H^{-1})T. \end{aligned} \quad (3.6)$$

Note that the Bayes estimator $\tilde{\beta}$ is a combination of the prior guess β_0 and the usual estimate $\hat{\beta}$. Note also that in the case of a flat prior, which corresponds to moving the elements of T so that its eigenvalues tend to infinity, then $G_T H^{-1} \doteq I - HT^{-1}$, in particular G_T tends to H , β tends to $\hat{\beta}$, and $\text{var}\{\beta \mid \text{data}\}$ becomes $\sigma^2 H$.

This is information of value, but the imminent interest is interpolation and its uncertainty. We find

$$\begin{pmatrix} Z(x) \\ Z_{\text{dat}} \end{pmatrix} \sim \mathcal{N}_{n+1} \left\{ \begin{pmatrix} f(x)' \beta_0 \\ F\beta_0 \end{pmatrix}, \sigma^2 \begin{pmatrix} K(x, x) + f(x)' Tf(x), & (k + FTf(x))' \\ k + FTf(x), & K + FTF' \end{pmatrix} \right\}.$$

Our Bayesian Kriger becomes

$$\begin{aligned} \hat{Z}_B(x) &= E\{Z(x) \mid \text{data}\} = f(x)' \beta_0 + (k + FTf(x))'(K + FTF')^{-1}(Z_{\text{dat}} - F\beta_0) \\ &= f(x)' \tilde{\beta} + k' K^{-1} (Z_{\text{dat}} - F\tilde{\beta}), \end{aligned} \quad (3.7)$$

with associated Bayesian prediction error

$$\begin{aligned} \sigma_{\text{be}}(x)^2 &= E\{\hat{Z}_B(x) - Z(x)^2 \mid \text{data}\} = \text{var}\{Z(x) \mid \text{data}\} \\ &= \sigma^2 [K(x, x) + f(x)' Tf(x) - (k + FTf(x))'(K + FTF')^{-1}(k + FTf(x))]. \end{aligned} \quad (3.8)$$

Again there is a natural correspondence for a flat prior, in that the Bayesian interpolator converges to $\hat{Z}(x)$ when the covariance matrix T for the β prior tends to infinity, and $\sigma_{\text{be}}(x)$ tends to $\sigma_{\text{pe}}(x)$. More informatively, calculations show that

$$\hat{Z}_B(x) = \hat{Z}(x) - (f(x) - F'K^{-1}k)'(I - G_T H^{-1})(\hat{\beta} - \beta_0),$$

and $G_T H^{-1}$ is close to $I - HT^{-1}$ when T is large compared to H . This means that in situations where T is large (vague prior information) and/or $\hat{\beta}$ is close to β_0 (good prior guess) the Bayesian and the traditional viewpoints lead to the same quantitative results, regarding interpolator and prediction variance, but with two quite different perspectives. Many statisticians have learned from experience that users very often prefer and relate better to the Bayesian interpretation.

The case of $T = 0$ corresponds to certainty about trend surface $f(x)'\beta = f(x)'\beta_0$, and $\hat{Z}_B(x) = f(x)'\beta_0 + k'K^{-1}(Z_{\text{dat}} - F\beta_0)$ is as in so-called simple Kriging with known trend. The other extreme is when T is very large, corresponding to prior ignorance about β , where $\hat{Z}_B(x)$ becomes as in universal Kriging. There is accordingly a “Bayesian bridge” from simple to universal Kriging.

The Bayesian apparatus becomes more useful and flexible when uncertainty about the scale factor σ^2 is modelled as well. The mathematically simplest Bayesian solution is the following: Suppose the prior distribution for (σ^2, β) is such that $\sigma^2 \sim \text{inverse gamma } (\alpha, \gamma)$ and $\beta | \sigma^2 \sim \mathcal{N}_p \{\beta_0, \sigma^2 T\}$, i.e.

$$(\sigma^2, \beta) \sim \text{const } (1/\sigma^2)^{\alpha+1} \exp(-\gamma/\sigma^2) \sigma^{-p} \exp\{-\frac{1}{2}(\beta - \beta_0)'T^{-1}(\beta - \beta_0)/\sigma^2\}.$$

Then calculations show that (σ^2, β) given data Z_{dat} is exactly of the same type, but with updated parameters:

$$(\sigma^2, \beta) | \text{data} \sim \text{inverse gamma } (\tilde{\alpha}, \tilde{\gamma}) \times \mathcal{N}_p \{\tilde{\beta}, \sigma^2 G_T\}.$$

Here $\tilde{\alpha} = \alpha + \frac{1}{2}n$ and $\tilde{\gamma} = \gamma + \frac{1}{2}n\hat{\sigma}^2 + \frac{1}{2}(\hat{\beta} - \beta_0)'L_T(\hat{\beta} - \beta_0)$, writing $\hat{\sigma}^2 = (Z_{\text{dat}} - F\hat{\beta})'K^{-1}(Z_{\text{dat}} - F\hat{\beta})/n$ for the non-Bayesian estimator and $L_T = T^{-1}G_T H^{-1}$. In particular the Bayesian predictor is as in (3.7) above, and the Bayes estimate of σ^2 is

$$\tilde{\sigma}^2 = E\{\sigma^2 | \text{data}\} = \frac{\alpha - 1}{\alpha - 1 + n/2} \sigma_0^2 + \frac{n/2}{\alpha - 1 + n/2} \{\hat{\sigma}^2 + (\hat{\beta} - \beta_0)'L_T(\hat{\beta} - \beta_0)/n\},$$

where $\sigma_0^2 = E\sigma^2 = \gamma/(\alpha - 1)$ is the prior guess value. The Bayesian prediction error is as in (3.8) but with $\tilde{\sigma}^2$ replacing σ^2 .

Once more the traditional methods emerge in the limiting case of vague Bayesian ignorance, which here corresponds to $\alpha \rightarrow 1$, $\gamma \rightarrow 0$, and $T \rightarrow \infty$.

3.1.4. Simulation

Suppose a simulated realisation $Z_s(\cdot)$ of a fully specified Gaussian random function is needed, perhaps conditioned on some observed values. First of all this means in practice that $Z_s(x)$ is to be simulated on some dense finite lattice of points only. Hence the problem seems simple, since textbooks on multivariate statistics give answers to such questions in terms of simple linear algebra. Such direct methods must involve matrix inversion, perhaps via Cholesky decomposition, and work well on good computers if the dimension of the problem is less than 3000. But in many spatial problems the grid net contains more than 10^6 points, and in such cases smarter simulation algorithms are called for.

One general approach to conditional simulation of the residual process $Z_r(x) = Z(x) - m(x, \beta)$, with associated residual observations $Z_r(x_i) = Z(x_i) - m(x_i, \beta)$ for $i = 1, \dots, n$, is

by decomposition. Let $Z_r(x) = \hat{Z}_r(x) + \{Z_0(x) - \hat{Z}_0(x)\}$, with $\hat{Z}_r(\cdot)$ the simple Kriging predictor based on the observations, $Z_0(\cdot)$ an arbitrary random function with characteristics identical to $Z_r(x)$, and $\hat{Z}_0(x)$ the simple Kriging predictor based on $Z_0(x_i)$ -data. From the orthogonality of $Z_r(\cdot)$ and $\hat{Z}_r(\cdot)$ it is easy to see that the decomposition is correct. Using this trick it suffices to simulate a random function with zero mean and covariance function $K(x, y)$ unconditionally.

The "turning band" procedure of Matheron (1973) provides a general tool for simulating random functions with a specified covariance structure. Several procedures working in the spectral domain are also available, see Ripley (1987, ch. 4).

Spatial and spatial-temporal problems often require simulation on extremely large grid nets, for which the procedures mentioned above will be too slow, even on very fast machines. In Omre *et al.* (1993) some very fast procedures for simulating random functions are presented and evaluated. The procedures use Markov characteristics valid in one dimension and generalise the idea to higher dimensions. The traditional procedures for simulation of fractal processes, see Feder (1988), are in the same spirit. The screening sequential procedure defined in Omre *et al.* (1993) is considerably more reliable than the traditional fractal procedures, however, albeit somewhat slower.

3.1.5. Estimating the covariance function

A point of some importance which has attracted renewed interest recently is that of the specification and estimation of the covariance function K . In the geostatistics tradition the K function has typically been regarded as "given and fixed" once "guessed at", which results in underestimation of interpolation variance, and, sometimes, suboptimal interpolation. Observe that the definition of and the interpretation of K and σ^2 in (3.2) are dependent upon the chosen regression surface model. If one adds a term to a previous regression trend, then K and σ^2 have changed meanings, and σ^2 will in fact tend to be smaller.

Consequences of using incorrect covariance functions have recently been studied by Watkins & Al-Boutiahi (1990) and Stein (1990a, b). Various methods for estimating parameters in such functions have been proposed, and a couple of these and a new one will be mentioned here.

Let us structure the problem somewhat and assume stationarity and isotropy, say $\sigma^2 K(x, y) = \sigma^2 R(\|x - y\|)$, identified by putting $R(0) = 1$, so that σ^2 is the variance of $Z(x)$ and $R(r)$ is the correlation between $Z(x)$ and $Z(y)$ whenever the locations are distance r apart. A non-parametric procedure which can be used with sufficient data is to group together all pairs of points with inter-distance in $(r - \delta, r + \delta)$, say, and estimate the covariance $R(r)$ based on these. This suggestion amounts to

$$\hat{\sigma}^2 \hat{R}(r) = \frac{\sum_{i,j} w(d_{ij}) \{z(x_i) - m(x_i, \hat{\beta})\} \{z(x_j) - m(x_j, \hat{\beta})\}}{\sum_{i,j} w(d_{ij})}, \quad (3.9)$$

where the sums are over all $n(n - 1)/2$ pairs, and $d_{ij} = \|x_i - x_j\|$ is the inter-distance between locations, but where $w(d)$ is 1 only for $d \in (r - \delta, r + \delta)$ and 0 outside. Somewhat more ambitiously a kernel type weighting of all pairs of data can be proposed. There are some difficulties with the direct use of estimators like (3.9), since one wants to ensure positive-definiteness of the covariance function to be used in Kriging. Therefore (3.9) type estimators are more often used as a means of deciding on some particular parameterised covariance function. Robustness issues are discussed in Omre (1984).

For parametric models $R(r) = R_\theta(r)$ there is an ongoing dispute over the merits of the mle. Maximising the likelihood over β gives $\hat{\beta}_\theta = (F' K_\theta^{-1} F)^{-1} F' K_\theta^{-1} Z_{\text{dat}}$, and then over σ gives $\hat{\sigma}_\theta^2 = Q_\theta/n$, where $Q_\theta = Q_\theta(\hat{\beta}_\theta) = (Z_{\text{dat}} - F\hat{\beta}_\theta)' K_\theta^{-1} (Z_{\text{dat}} - F\hat{\beta}_\theta)$. Here K_θ is the matrix of $R_\theta(d_{ij})$. The resulting profile likelihood must finally be maximised over θ . This amounts to minimising the rather intricate function

$$-\log L(\hat{\beta}_\theta, \hat{\sigma}_\theta, \theta) = \frac{1}{2}n \log(Q_\theta/n) + \frac{1}{2} \log \{\det(K_\theta)\} + \frac{1}{2}n. \quad (3.10)$$

There are first of all numerical problems associated with minimising this difficult and possibly multimodal function, and secondly it is not clear that using the maximum likelihood estimator should be a good choice *per se*. See the dispute of Mardia & Marshall (1984), Warnes & Ripley (1987), Ripley (1988, ch. 2), and Mardia & Watkins (1989). Other estimation methods have also been proposed, see Switzer (1984), Stein (1987), and Vecchia (1989). Pseudo-likelihood methods in the manner of Besag (1974) and Jensen & Møller (1991) can also be used.

Let us briefly describe yet another method, the maximum quasi-likelihood procedure of Hjort (1993). This is really a class of methods, and the simplest among them is the following: consider

$$\text{ql}(\beta, \sigma, \theta) = \prod_{i < j} g(z_i, z_j | d_{ij}, \beta, \sigma, \theta), \quad (3.11)$$

where the product is over all $N = n(n - 1)/2$ pairs of distinct observations, and the g -term is the model-given probability density of $(Z(x_i), Z(x_j))$ for points lying distance d_{ij} apart. Maximising this is easily carried out, as indicated below, without numerical problems. Why is maximising ql a sensible procedure? Suppose $g(z, z' | r)$ is the *true* probability density for a pair $(Z(x), Z(x'))$ with inter-distance r . Divide the distance range $[0, \infty)$ into small intervals, and sort for each interval together those pairs in $\log \text{ql}$ that have inter-distance d_{ij} close to the corresponding distance, say r . An ergodic argument shows that

$$N^{-1} \log \text{ql}(\beta, \sigma, \theta) \doteq \int_0^\infty \left[\int \int g(z, z' | r) \log g(z, z' | r, \beta, \sigma, \theta) dz dz' \right] H(dr),$$

in which $H(dr)$ is the distribution of the distance $\|x - x'\|$ between a randomly drawn pair of points. Hence maximising ql aims at finding the parameter values that minimise the particular distance function

$$\Delta[g(\cdot, \cdot | \cdot), g(\cdot, \cdot | \cdot, \beta, \sigma, \theta)] = \int_0^\infty \Delta_r[g(\cdot, \cdot | r), g(\cdot, \cdot | r, \beta, \sigma, \theta)] dH(r),$$

in which the inner distance function between the true density g_r and the modelled density g_r^* for $(Z(x), Z(x'))$ is the Kullback–Leibler one, $\int \int g_r \log \{g_r/g_r^*\} dz dz'$. In particular the proposed maximum ql method leads to consistent parameter estimates under suitable mild regularity conditions, but we avoid here the precise definition of the asymptotic framework.

Let us see how this works out in the three-parameter model in which $Z(x) \sim \mathcal{N}\{\beta, \sigma^2\}$ and the covariance is $\sigma^2 R_\theta(r)$ for points lying distance r apart. As with the ordinary likelihood method the programme is to maximise w.r.t. β , then over σ , and finally over θ . This leads to the following, where we take the liberty of using the same notation as for the ml case above: let first $\hat{\beta}_\theta$ minimise

$$Q_\theta(\beta) = \sum_{i < j} \frac{1}{2} \frac{(z_i - \beta)^2 + (z_j - \beta)^2 - 2R_\theta(d_{ij})(z_i - \beta)(z_j - \beta)}{1 - R_\theta(d_{ij})^2},$$

indeed

$$\hat{\beta}_\theta = \frac{\sum_{i < j} \frac{(z_i + z_j)/2}{1 + R_\theta(d_{ij})}}{\sum_{i < j} \frac{1}{1 + R_\theta(d_{ij})}}$$

Then let $\hat{\sigma}_\theta^2 = Q_\theta(\hat{\beta}_\theta)/N = Q_\theta/N$. The remaining task is to minimise

$$-\log \text{ql}(\hat{\beta}_\theta, \hat{\sigma}_\theta, \theta) = N \log(Q_\theta/N) + \sum_{i < j} \frac{1}{2} \log \{1 - R_\theta(d_{ij})^2\} + N \quad (3.12)$$

w.r.t. θ , and this is much simpler than minimising (3.10).

There is a connection to the simple non-parametric correlation function estimator (3.9), in that if the particular model which postulates piece-wise constant $R(\cdot)$ is used, then the ql-solution can be shown to be close to (3.9), and the related quasi-likelihood which only uses data pairs with approximate distance r comes even closer.

The virtues of the ql method are that the numerical maximisation problem is much simpler than for the ordinary likelihood method, its relatively wide applicability, that its behaviour is better understood, and that it in fact behaves well. It does not claim to be optimal, and in the asymptotic framework where the region expands to produce nearly independent copies the ml is better. This framework is somewhat inappropriate, however.

Instead of using all pairs one could take the product over all $g(z_i, z_{n(i)} | d_i, \beta, \sigma, \theta)$ terms, where $z_{n(i)}$ is nearest neighbour to z_i , with interdistance d_i . Reasoning similar to above shows that this estimation procedure aims at minimising a different distance criterion from true model to parametric approximand, using the $H_0(dr)$ distribution for nearest neighbour distances instead of $H(dr)$ for an arbitrary distance. And there are several related alternative methods. The two procedures above care only about data-pairwise aspects of the model, and would not necessarily be good enough for prediction purposes, for example. One could with some additional efforts use three data points at a time, to fit the empirical three-point-sets distribution to the model. A fuller account is given in Hjort (1993).

The discussion above concerns the basic model of 3.1.1. More sophisticated models are sometimes used when there is enough data, and then appropriate generalisations of the covariance function estimators above are needed. In the growing area of environmental monitoring, for example, there are often abundant time series data in fixed geographical locations, which invites better ways of modelling residual variation. Covariance function modelling and estimation in such frameworks are discussed in Sampson & Guttorp (1992), Loader & Switzer (1992), and Høst *et al.* (1994). See section 4.4.

3.2. Mosaic processes

Here we describe processes that divide the reference space into a set of disjoint segments and assign a label to each segment. The Markov random fields (mrf), defined on regular lattices, are enjoying increasing popularity, though mixed with a widened understanding of the inherent limits and difficulties of the approach. The fundamental simulation scheme called the Metropolis algorithm is explained. Some recently developed amendments and alternative models, including constrained mrfs and semi-mrfs are then discussed. Finally a couple of non-lattice situations are considered.

3.2.1. Markov random fields

We are concerned with stochastic models for the distribution of classes on lattices. One class of such models is the class of mrf's. The following is a brief description of mrf's and some of the statistical properties of such models. We discuss how to estimate parameters of such models, based on an observed "true scene", and describe methods to simulate realisations of mrf's. One needs to be able to simulate both unconditionally and data-conditionally from a specified mrf. Global constraints that one sometimes needs to be able to impose on simulated scenes include preservation of information in some locations and the desire to keep frequencies of some or all classes near specified levels.

For a given lattice system of sites we first need to introduce the notion of a clique. Because of the strong association to image analysis we shall mainly think of the sites as picture elements, or pixels. Assume a system of neighbourhoods has been defined. Then define a clique Q as a set of pixels all of which are neighbours of each other. Note the sociological appropriateness of the term. If "neighbours" means nearest neighbours, not including the diagonal ones, then all cliques are of the type

$$\{(i, j), (i + 1, j)\} = * * \quad \text{or} \quad \{(i, j), (i, j + 1)\} = * * .$$

If also diagonal neighbours are included, so that each site has eight neighbours, then there are nine types of cliques:

$$\begin{array}{cccccccccc} * & * & * & * & * * & * & * * & * & * * \\ * * & * & * & * & * & * & * * & * & * * \end{array} \quad (3.13)$$

The mrf class of probability distributions, or Gibbs processes, are those that satisfy

$$p(\mathbf{x}) = p(x_1, \dots, x_N) = \text{const} \exp \left[\sum_{i=1}^N \alpha_i(x_i) + \sum_Q V_Q(x_Q) \right]. \quad (3.14)$$

Here $\mathbf{x} = (x_1, \dots, x_N)$ is a long vector of class labels, say among $\{1, \dots, K\}$, in N sites or pixel locations; $V_Q(x_Q) = V_Q(x_{Q,1}, \dots, x_{Q,m(Q)})$ is the "potential" associated with clique Q ; $\alpha_i(1), \dots, \alpha_i(K)$ are class and position dependent parameters that can be tied to the prior probabilities for the various classes; and the sum is over all cliques. One important consequence is that

$$p_i(k \mid \text{rest}) = \Pr \{ \text{class} = k \text{ in pixel } i \mid \text{rest} \} = \frac{\exp \{ \alpha_i(k) + A_i(k, x_{\partial i}) \}}{\sum_{l=1}^K \exp \{ \alpha_i(l) + A_i(l, x_{\partial i}) \}}, \quad (3.15)$$

in which $x_{\partial i}$ is the collection of classes in the neighbouring pixels lying around pixel i . In fact $A_i(k, x_{\partial i}) = \sum_{Q: i \in Q} V_Q(x_Q; x_i = k)$. This means in particular that $p_i(k \mid \text{rest})$ depends upon only $x_{\partial i}$, i.e.

$$p_i(k \mid \text{rest}) = p_i(k \mid x_{\partial i}). \quad (3.16)$$

This is the Markov property. The probabilities (3.16) are called the local characteristics of the mrf. One could call $A_i(k, x_{\partial i})$ the award function for window $\{i\} \cup \partial i$ around pixel i . The model encourages realisations with high awards.

The remarkable Hammersley–Clifford–Besag theorem identifies processes having the Markov property (3.16) with those having the Gibbs property (3.14), under a positivity condition; see Besag (1974). When faced with the task of constructing a suitable mrf to describe a certain phenomenon it is usually simpler to think "local Markov" in terms of award functions and local characteristics than thinking "global Gibbs" in terms of the potentials.

The following mrf is among the structurally simplest, but has proved useful in image restoration (see section 3.4.3 below) and in other areas: use 3×3 neighbourhoods, and cliques of size two only in (3.13), use class-dependent but position-independent $\alpha_i(x_i) = \alpha(x_i)$, and encourage spatial continuity by putting $V_{(i,j)}(x_i, x_j) = \beta I\{x_i = x_j\}$ in (3.14). This spatial mosaic model has $K + 1$ parameters and local characteristics

$$p_i(k \mid \text{rest}) = \frac{\exp \{ \alpha(k) + \beta H_i(k, x_{\partial i}) \}}{\sum_{l=1}^K \exp \{ \alpha(l) + \beta H_i(l, x_{\partial i}) \}}, \quad (3.17)$$

in which $H_i(x_i, x_{\partial i})$ is the number of x_i 's neighbours that agree with it. The award is a value in $0, \beta, 2\beta, \dots, 8\beta$.

The mrfs obviously form a very wide class. Including other cliques carefully chosen from larger neighbourhoods and perhaps finer parameterisation than a crude β for each gives one the possibility of including various types of prior knowledge about local structure into the model. In application 2.3.2 described in section 2 we fitted mrfs with $K = 12$ classes that came in four subgroups and 5×5 -windows with up to 16 different cliques and up to six different β -parameters, see Hjort *et al.* (1989). In application 2.5.2 similarly complex mrfs have been used to model handwritten numbers, see Prupp (1990). See also the general discussion by Ripley (1992).

Remark 9. It is instructive to compare the 2-D Markov property to the corresponding 1-D one, i.e. for Markov chains $\{x_n\}$. The classical definition of Markov-ness in the chain case is that the distribution of some x_n given the complete past depends on x_{n-1} only. The alternative characterisation that the distribution of x_n given both past and future depends on the nearest neighbours only lends itself much more naturally to higher dimensions. \square

3.2.2. Parameter estimation from a single scene

Consider for concreteness a mrf with local characteristics of the form (3.17), but with a more general award function structure

$$A_i(x_i, x_{\partial i}) = \beta_1 H_i^{(1)}(x_i, x_{\partial i}) + \dots + \beta_p H_i^{(p)}(x_i, x_{\partial i}) = \beta' H_i(x_i, x_{\partial i}),$$

for certain parameters β_1, \dots, β_p and certain simple functions $H_i^{(j)}(x_i, x_{\partial i})$. The task is to obtain estimates for $\alpha(k)$ s and β s (and perhaps further parameters, see section 3.4 below), from a single realisation \mathbf{x} of the assumed mrf process. The $\alpha(k)$ s are tied to the prior probabilities in pixel i , but in a rather involved way, because of interaction with the β s.

This is in one way a simply structured exponential model of classical form, say $p(\mathbf{x}) = c(\beta) \exp(\beta' V(\mathbf{x}))$ (having subsumed $\alpha(k)$ s in new β s). The maximum likelihood method is seen to be equivalent to solving $V^{(j)}(\mathbf{x}) = \mu^{(j)}(\beta_1, \dots, \beta_p)$ for $j = 1, \dots, p$, where $\mu^{(j)}(\beta) = -\partial \log c(\beta)/\partial \beta_j$ is the expected value of $V^{(j)}(\mathbf{x})$ under the model. These equations cannot be solved directly, due to the formidable normalisation constant $c(\beta)$, defined as a sum of K^N terms. This is the ‘partition function’ of statistical mechanics, and to give an idea of its complexity it suffices to mention that a Nobel Prize was awarded Nils Onsager for just providing an approximation. But ml estimates can be computed after all, through the use of extensive simulations, see the idea sketched in Künsch (1986). Pickard (1987) managed the simplest case of two equally likely classes and a single β parameter. See also Gidas (1991) for a precise and general account.

The alternative maximum pseudo-likelihood method is much simpler to implement, and has become the method of popular choice. It consists of maximising $\text{pl} = \prod_{i=1}^N p_i(x_i | x_{\partial i})$ w.r.t. the parameters of the model. In the case considered

$$\log \text{pl} = \sum_{i=1}^N \left[\alpha(x_i) + \beta' H_i(x_i, x_{\partial i}) - \log \left(\sum_{l=1}^K \exp \{\alpha(l) + \beta' H_i(l, x_{\partial i})\} \right) \right], \quad (3.18)$$

and this function can be maximised numerically w.r.t. the parameters β_1, \dots, β_p and the parameters of $\alpha(k)$. To give expressions for partial derivatives, let again $\alpha_i(k)$ terms be subsumed into the H_i functions to form a larger V_i vector, and let $E_i V_i(\cdot, x_{\partial i})$ denote the mean of $V_i(X_i, x_{\partial i})$ conditional on $x_{\partial i}$, so that $X_i = k$ with probability $p_i(k | x_{\partial i})$, and let $\text{var}_i V_i(\cdot, x_{\partial i})$ be the accompanying variance matrix. Then

$$\frac{\partial \log \text{pl}}{\partial \beta_j} = \sum_{i=1}^N \{V_i^{(j)}(x_i, x_{\partial i}) - E_i V_i^{(j)}(\cdot, x_{\partial i})\},$$

$$\frac{\partial^2 \log \text{pl}}{\partial \beta_j \partial \beta_m} = - \sum_{i=1}^N \{\text{var}_i V_i(\cdot, x_{\partial i})\}_{j,m},$$

which shows that $\log \text{pl}$ is concave and well-behaved as a function of β .

A natural first step is to employ parameter-free prior probabilities $\exp \{\alpha(k)\} = \pi(k)$, for example position-independent ones and taken from some “prior scene”, so that $\log \text{pl}$ needs to be maximised only w.r.t. β_1, \dots, β_p . The maximisation could if necessary go on in an iterative manner. If $\alpha(k)$ is treated as an unknown parameter then $\partial \log \text{pl} / \partial \alpha(k) = 0$ leads to the natural equation $\pi_k(\mathbf{x}) = (1/N) \sum_{i=1}^N I\{x_i = k\} = (1/N) \sum_{i=1}^N p_i(k | x_{\partial i})$. In application 2.3.2 we have also had occasion to use and fit mrfs with non-linear exponents. All in all parameter estimation using maximum pseudo-likelihood requires some reliable maximisation algorithms, along with a flexibly structured environment to handle such data structures, but is not a major obstacle.

Remark 10. Why does pseudo-likelihood work? And is it clear that ordinary maximum likelihood works, if arduously carried through? Again it is instructive to consider the one-dimensional case of Markov chains. Suppose data $\{x_a : a = 0, \dots, n\}$ are observed from some stationary process and are to be fitted to some parametric first-order Markov chain model $\Pr\{X_{a+1} = j | X_a = i\} = p_\beta(j | i)$. Let $p_\beta(i) = \Pr\{X_a = i\}$ be the accompanying marginal distribution, which is the equilibrium distribution. The ml procedure and the pl procedure maximise respectively

$$L(\beta) = \prod_{a=1}^n p_\beta(x_a | x_{a-1}) = \prod_{i,j} p_\beta(j | i)^{N(i,j)}$$

and

$$\text{pl}(\beta) = \prod_{a=1}^{N-1} p_\beta(x_a | x_{a-1}, x_{a+1}) = \prod_{i,j,k} p_\beta(j | i, k)^{N(i,j,k)}$$

in self-explanatory notation.

To examine the aims of ml and pl let us merely postulate that the true underlying model for the chain is some stationary distribution with $\Pr\{X_a = i, X_{a+1} = j\} = p(i, j) = p(i)p(j | i)$ and $\Pr\{X_a = i, X_{a+1} = j, X_{a+2} = k\} = p(i, j, k)$. By ergodic arguments $N(i, j)/n$ and $N(i, j, k)/n$ tend to $p(i, j)$ and $p(i, j, k)$ in probability. Hence

$$\frac{1}{n} \log L(\beta) \xrightarrow{P} \lambda_1(\beta) = \sum_{i,j} p(i, j) \log p_\beta(j | i) = \sum_i p(i) \sum_j p(j | i) \log p_\beta(j | i)$$

and

$$\begin{aligned} \frac{1}{n} \log \text{pl}(\beta) &\xrightarrow{p} \lambda_{\text{pl}}(\beta) = \sum_{i,j,k} p(i,j,k) \log p_\beta(j | i, k) \\ &= \sum_{i,k} p(i, \cdot, k) \sum_j p(j | i, k) \log p_\beta(j | i, k), \end{aligned}$$

in which $p(i, \cdot, k)$ is the sum of $p(i, j, k)$ over all j . It follows that ml and pl in general aim at different “least false” or “most appropriate” parameter values. The ml procedure aims at and will be consistent for the parameter value β_0 which is least false according to the distance measure

$$\text{dist}_1 \{ \text{truth, model} \} = \sum_i p(i) \sum_j p(j | i) \log \frac{p(j | i)}{p_\beta(j | i)},$$

a weighted sum of Kullback–Leibler distances between true and modelled transition probabilities. The pl procedure, on the other hand, aims at and is consistent for the second parameter value β_1 that minimises the different distance measure

$$\text{dist}_{\text{pl}} \{ \text{truth, model} \} = \sum_{i,k} p(i, \cdot, k) \sum_j p(j | i, k) \log \frac{p(j | i, k)}{p_\beta(j | i, k)},$$

another weighted sum of Kullback–Leibler distances, this time between true and modelled local characteristics $\Pr \{ X_a = j | X_{a-1} = i, X_{a+1} = k \}$. One may also study limit distributions in this framework. The ml is somewhat better on the model’s home turf.

This discussion is pertinent considering our proclaimed view that models should be fitted and “adapted” but not necessarily trusted. \square

3.2.3. Unconditional simulation

The task considered is that of simulating realisations of a specified mrf. This amounts to simulating from a discrete probability distribution $p(\mathbf{x})$, see (3.14), on an enormous but finite space, the set S of all K^N possible combinations of classes on the given lattice. In application 2.3.2 mentioned in section 2 we worked with $K = 12$ classes and for example $N = 200 \times 100$ sites or pixels in the scene, which leads to enormous numbers of size $10^{20,000}$ and the like for the number of different scenes. Numbers become mind-boggling in 3-D! Ordinarily methods can of course not cope with this kind of magnitude of the problem.

The simulation tricks to be used employ (MC²) Markov Chain Monte Carlo methods. Assume that a huge transition matrix for a Markov chain with state space S has elements $m(\mathbf{x}, \mathbf{x}')$ that satisfy the reversibility criterion

$$p(\mathbf{x})m(\mathbf{x}, \mathbf{x}') = p(\mathbf{x}')m(\mathbf{x}', \mathbf{x}) \quad (3.19)$$

for all possible scenes, in which $p(\mathbf{x})$ is the mrf distribution given in (3.14). Then one can show that $p(\cdot)$ is the equilibrium distribution for the Markov chain. To sample from $p(\cdot)$, therefore, one might choose a convenient transition matrix with elements $m(\cdot, \cdot)$ obeying (3.19), and then run it until equilibrium seems to have settled in.

There are several possibilities of choice of $m(\cdot, \cdot)$. A convenient class of such can be described as follows. Choose first a symmetric transition matrix with elements $q(\mathbf{x}, \mathbf{x}')$ that are mostly equal to zero. Assume the simulation chain has come to scene $\mathbf{x} = (x_1, \dots, x_N)$. Select a potential new scene $\mathbf{x}'(x'_1, \dots, x'_N)$ from $q(\mathbf{x}, \cdot)$, and move from \mathbf{x} to \mathbf{x}' with

probability $\min \{1, p(\mathbf{x}')/p(\mathbf{x})\}$, otherwise remain at \mathbf{x} . This is the Metropolis algorithm, or perhaps “the class of Metropolis algorithms”. One can check that (3.19) holds. The basic idea behind this simulation trick seems to be due to Metropolis (*see* Metropolis *et al.* (1953), and Hastings (1970) and Ripley (1987, ch. 4) for more statistical accounts).

It remains to specify $q(\cdot, \cdot)$, and again several options are available. One possibility is to have $q(\mathbf{x}, \mathbf{x}')$ positive only if \mathbf{x} and \mathbf{x}' differ at a single site. If this site is i , and $\mathbf{x} = (x_1, \dots, k, \dots, x_N)$ and $\mathbf{x}' = (x_1, \dots, l, \dots, x_N)$, then

$$\frac{p(\mathbf{x}')}{p(\mathbf{x})} = \frac{p_i(l \mid \text{rest})}{p_i(k \mid \text{rest})} = \frac{\exp \{\alpha_i(l) + A_i(l, x_{\hat{o}_i})\}}{\exp \{\alpha_i(k) + A_i(k, x_{\hat{o}_i})\}},$$

see (3.15). One feasible simulation method, for generating a single realisation \mathbf{x} , is therefore as follows. Start out with some initial scene, for example with class labels simulated independently in different pixels, from the prior probabilities $\pi_i(k)$. Carry out complete iteration cycles until apparent equilibrium, where one iteration cycle means a full scan over the scene. And when such a scan visits site i , choose class label l randomly, and let x_i change from its current label k to l with probability $\min \{1, p_i(l \mid \text{rest})/p_i(k \mid \text{rest})\}$.

Among several other convenient methods the so-called Gibbs sampler is perhaps the most popular. It consists of running complete iteration cycles until convergence, as above, with the following schedule for changing class labels during a full scan. If the current t th generation scene is $\mathbf{x}_t = (x_{1,t}, \dots, x_{N,t})$, choose a random class label $x_{i,t+1} = k$ for site i according to the local probabilities $p_i(k \mid \mathbf{x}_t - \{i\}) = p_i(k \mid \text{rest}_t)$. Only empirical evidence can be given for preferring one simulation scheme to another. The “current folklore”, in this hectic but still young field of stochastic simulation, seems to favour the Gibbs sampler. Some references are Geman & Geman (1984), Gidas (1985), Ripley (1987), and Tjelmeland & Holden (1993).

3.2.4. Constrained simulation

A pleasing facet of the mrf simulation scheme is that one can condition on known class values in some locations. Just go on running the Markov simulation chain as in the previous subsection, for example the Gibbs sampler, but with the class labels fixed at their known values, at all sites where such labels are known.

In some situations realisations of the mrf are only close to reality if the areas covered by each class are somewhat close to prior conceptions. It is therefore important to be able to constrain simulated realisations of a mrf to have class proportions equal to or close to specified values. This is not an easy problem, and some confusion surrounds the few methods that have been proposed in the literature. One possibility is a spin exchange method due to Flinn, *see* Hjort *et al.* (1989) for a brief review. Another and in many ways more promising avenue is to consider a new stochastic model with probability distribution

$$p_\sigma(\mathbf{x}) = p_\sigma(x_1, \dots, x_N) = \text{const } p(x_1, \dots, x_N) \exp \{-\sigma \Delta(x_1, \dots, x_N)\}, \quad (3.20)$$

in which $\Delta(x_1, \dots, x_N) = \sum_{k=1}^K (N_k/N - \pi_k^0)^2$ is a measure of discrepancy between the class frequencies of the scene \mathbf{x} and the specified class frequencies. It is for each given σ possible to simulate from $p_\sigma(\cdot)$, by methods similar to those outlined above, even though this random field is not any longer a mrf w.r.t. the neighbourhood system. The idea is then to let σ slowly increase as the Markov simulation chain moves on. The result is that the observed class proportions are forced towards the prescribed π_k^0 s.

Let us elaborate this point. The local characteristics of the new model are of the form

$$\begin{aligned} p_\sigma(x_i = k \mid \text{rest}) &= \frac{p_\sigma(x_1, \dots, k, \dots, x_N)}{\sum_{l=1}^K p_\sigma(x_1, \dots, l, \dots, x_N)} \\ &= \frac{\exp [\alpha_i(k) + A_i(k, x_{\hat{o}}) - \sigma \Delta(x_1, \dots, k, \dots, x_N)]}{\sum_{l=1}^K \exp [\alpha_i(l) + A_i(l, x_{\hat{o}}) - \sigma \Delta(x_1, \dots, l, \dots, x_N)]}. \end{aligned}$$

This expression is generally valid for any given Δ -function of discrepancy between observed and ideal characteristics of the scene $\mathbf{x} = (x_1, \dots, x_N)$. With the present Δ -function further simplification is possible, see Hjort *et al.* (1989), and it is not difficult to use the Gibbs sampler, as follows. Complete iteration cycles of simulated scenes are run until apparent equilibrium, where one cycle is a full scan over the scene. During one such scan, suppose the current scene is $\mathbf{x} = (x_1, \dots, k_0, \dots, x_N)$. Then move from k_0 to k with probability $p_\sigma(x_i = k \mid \text{rest})$.

Remark 11. This approach is useful also for other $\Delta(\mathbf{x})$ -measures of discrepancy between observed and specified characteristics of the scene. It has been used in various forms in projects we have worked on, also for marked point processes (see sections 3.3 and 4.3). It was also proposed in a remark by Green (1986) in the context of mrf's for image restoration. \square

3.2.5. Semi-mrf and other mosaic processes

Yes, there are others, and the mrf's have perhaps had too much attention during the last few years. Let us here briefly mention some other approaches for modelling mosaic type phenomena.

There are attempts at making the ordinary mrf's less crucially dependent on the given grid, or given resolution. See Nicholls & Petrou (1993) for ideas and experiments on renormalisation group transformation methods. In some applications the boundaries between class patches are approximately linear, and the resulting mosaic image is polygonal. A natural approach is therefore to model the boundaries themselves, and perhaps model class labels separately afterwards. Switzer (1965) gives a class of Poisson line models, which has been used by Owen (1984) and Hjort (1985a) in image reconstruction problems, cf. section 3.4.2 below. The Switzer process has a Markov property along line transects, but is not spatially Markovian; the distribution of lines and class labels in the interior of a convex bounded region given what is outside the region is not determined by knowledge of lines and class labels on the boundary. A more complex process which has this spatial Markov property has been introduced by Arak & Surgailis (1989). Clifford & Middleton (1989) outline mathematical properties of its potential use in image reconstruction, where there are problems of simulating from the posterior distribution. See also Arak *et al.* (1994). There is also a differently motivated approach based on coverage processes, covered by Hall (1988). Yet another method involves tessellations, Voronoi cells, etc., see Ripley (1981, ch. 4).

Let us finally describe a semi-mrf approach that aims at combining some of the convenient features of the mrf methods with some larger-scale modifications. A realisation \mathbf{x} on a lattice defines "bodies", maximally connected sets of sites that belong to the same class. There are several possibilities for making this intuitive notion mathematically precise; see Holden & Tjelmeland (1990) for one definition that is even valid in 3-D. The idea is that users often know something about the typical sizes and forms and directions of bodies, at least for some

of the classes. If bodies of sites from class 2 tend to be elliptical, for example, one can try to specify distributions of the angle and the axis lengths, and then build a model that is locally a mrf but encourages such bodies to take place, using a more complicated version of (3.20). One would also want to be able to gear such a model towards respecting certain borders on the scene, or towards respecting prior information of the type “bodies from class 3 very rarely touch bodies from class 4”. The challenge is to build a sufficiently general and flexible model that can do this and which makes it possible to simulate realisations. This rather ambitious scheme is carried out in Holden & Tjelmeland (1990) and in Tjelmeland & Holden (1993), using models of the form

$$p(\mathbf{x}) = \text{const } p_0(\mathbf{x}) \exp \left\{ \sum_{b=1}^B \gamma_b W_{\text{class}(B_b)}(B_b, f(B_b)) - \sum_{d=1}^D \sigma_d \Delta_d(\mathbf{x}) \right\},$$

where $p_0(\mathbf{x})$ is some ordinary mrf, perhaps with 5×5 windows, the B_b s are bodies with features $f(B_b)$ such as size and form, the W s are award functions that encourage bodies to meet certain specifications, and the Δ_d s are certain discrepancy functions. This model has been used to build a programme system that produces simulations of reservoir architecture and reservoir properties for oil companies; see the references mentioned.

3.3. Event processes

Here we briefly review some theory for spatial point processes. To solve problems associated with applications described in sections 2.3.3, 2.3.4 and 2.4.2, for example, one typically needs to model both location of points and associated “marks”. This is one of the combinations of processes we treat in section 3.4. It will be seen that natural and simple models sometimes are easy to construct, but that parameter estimation and model verification typically become difficult tasks. Simulation based inference and use becomes important.

3.3.1. Spatial point processes

How can we model the geographical positions of a collection of points, or small objects, in a given region? The simplest possibility is a Poisson process with constant intensity λ , say, which postulates that the number of points falling in disjoint regions are independent and Poisson distributed with parameters equal to λ times the areas of the regions. A wider class of models emerges by allowing the intensity to vary, perhaps even stochastically. Many important phenomena are not well modelled by even such varying-intensity or doubly stochastic Poisson processes, however. A typical feature is that the smallest inter-point distances are not as small as they would be under Poisson-ness.

A rich class of models for point patterns is the pairwise interaction processes, with density (Radon–Nikodym derivative) $f(\{x_1, \dots, x_n\}) = \text{const } \prod_{i=1}^n g(x_i) \prod_{i < j} h(\|x_i - x_j\|)$ w.r.t. the Poisson point process (with unit intensity, say). Here $g(x)$ is some non-negative function of geographical position, often taken to be constant; $h(r)$ is some non-negative function of the distance r between points, usually bounded by 1; and the integration constant is unwieldy. One is usually content to study models for fixed number of points n , where the pairwise interaction model says

$$f_n(x_1, \dots, x_n) = \text{const } \prod_{i=1}^n g(x_i) \prod_{i < j} h(\|x_i - x_j\|).$$

There are some measure-theoretic details to work through here; see Lotwick & Silverman (1981) and Baddeley & Møller (1989) for good accounts. Typical examples include

$$h(r) = \begin{cases} \gamma & \text{if } r < \rho, \\ 1 & \text{if } r \geq \rho, \end{cases}$$

which gives the so-called Strauss model, and other simple functions that in some way climb from 0 (or from some extra parameter ε) to 1 when r runs from 0 to ρ . The special case $\gamma = 0$ is possible in the Strauss model, and corresponds to a Poisson process which is never allowed to have points closer to each other than ρ ; this is also called a hard core process with hard core distance ρ .

With these choices for $h(\cdot)$ there is an upper limit after which no interaction occurs, and the product above is only over all neighbour pairs, where being neighbours means $\|x_i - x_j\| < \rho$. This invites Markov connections. Indeed the random collection $X = \{x_1, \dots, x_n\}$ has a spatial Markov property, see Baddeley & Møller (1989). General references for modelling spatial point patterns include Ripley (1977, 1981, 1988, 1989a, b) and Stoyan *et al.* (1987). Some geological applications are described in Omre (1992).

3.3.2. How to simulate

An operational definition of “understanding a model” is that one should be able to simulate realisations from it. For the Strauss model with $\gamma = 0$, for example, one could conceivably generate points from the Poisson model (which for fixed n means simulating from the uniform distribution in the region considered) and only keep those realisations that have smallest distance at least ρ . This can be seriously inefficient, and better methods are called for.

What seems now to be the best way is that of the spatial birth-and-death processes introduced for this task by Ripley (1977) and Ripley & Kelly (1977). The scheme is (usually) to delete one of the n points at random and then add back another one, with probability

$$\Pr \{ \text{add } x_n \mid x_1, \dots, x_{n-1} \} = \frac{f_n(x_1, \dots, x_n)}{f_{n-1}(x_1, \dots, x_{n-1})} = g(x_n) \prod_{i=1}^{n-1} h(\|x_i - x_n\|)$$

assigned to position x_n . The point is that this birth-and-death process has a unique equilibrium distribution which is exactly the point process with density f . Note that the probability above only depends on the new point and its neighbours, in case of a h -function with finite range. The simulation is usually carried out using some appropriate rejection sampling scheme. Cleverness in doing this is often essential for the algorithm to work fast enough.

3.3.3. How to read a point pattern

There is a need to summarise the main features of a given picture of point locations. One can think of a “first order” summary picture, where the intensity of points per unit area is estimated in some smoothing fashion. This part is related to the $g(x)$ -function of section 3.3.1. When the intensity can reasonably be assumed to be homogeneous over a region one needs a suitable “second order” summary. The method of choice here has become Ripley’s K or L functions. The $K(t)$ function is a measure related to the covariance between the number of points falling in two areas, and under ideal Poisson conditions $K(t) = \pi t^2$. This is the motivation for studying $L(t) = \{K(t)/\pi\}^{1/2}$ instead. We refer to Ripley (1981, 1988) and to Stoyan *et al.* (1987) for proper definitions and constructions of estimators that in various ways take edge effects into account. These summary curves can be used to detect non-Poisson behaviour, to suggest other models, and to estimate parameters in such.

If the “points” have associated areas, for example, then other summary characteristics are needed as well. Ripley (1986; 1988, ch. 6) surveys several such based on morphology and

Serra calculus. See also 3.4.8 below for more general processes that combine point locations with “marks”.

3.3.4. Estimating parameters

Even a simple-looking model like the Strauss model with $g = 1$ is notoriously difficult to estimate from data. Maximum likelihood estimation is difficult because of the intractable integration constant, but can be carried out through simulation procedures. One Monte Carlo method is described in Ripley (1988, ch. 4), and the rudiments of a general stochastic approximation method are presented in Moyeed & Baddeley (1991), following an idea of Künsch. Versions of pseudo-likelihood methods are studied in Grabarnik & Särkkä (1992), Särkkä (1993), and in Jensen & Møller (1991). Another method via conditioning of Palm probabilities has been developed by Takacs & Fiksel; see Fiksel (1988) and Grabarnik & Särkkä (1992), who point out connections to pseudo-likelihood again. Non-parametric estimation of the $h(r)$ function is very difficult to do with reasonable precision, and is perhaps only feasible with very large sample sizes. It can nevertheless be carried out, see Diggle *et al.* (1987), and be used as a data analytic summary. Goodness of fit is considered in Diggle (1979).

3.4. Combinations

In this final subsection a couple of naturally occurring cross-situations are discussed, in which Gaussian random functions, mosaic processes, and event processes appear in combination.

3.4.1. Hidden Markov fields and image restoration

The following situation occurs naturally in image analysis applications. There is an underlying true image \mathbf{x} with value x_i in pixel i , but corrupting noise is present and $y_i = x_i + \varepsilon_i$ is observed instead. More generally there could be a vector y_i carrying information about the true x_i , for example in the form of $y_i | x_i \sim \mathcal{N}_d\{\mu(x_i), \Sigma(x_i)\}$.

In some situations there is a low number of possible values for x_i , say $1, \dots, K$, which suggests using mosaic models of the type discussed in 3.2 for this “hidden truth”. Thus the land cover classes in a remote sensing application could be modelled as a mrf with an appropriate neighbourhood structure. In other situations there is a larger number of possible x_i -values, and the \mathbf{x} process could be viewed as a discretisation of a continuous random function. The noise is typically assumed to be Gaussian and independent from pixel to pixel given the \mathbf{x} labels. In applications we have worked with involving fine-resolution multi-channel satellite data, the measurements from the same underlying class have indeed been quite Gaussian, but have exhibited strong autocorrelation. In many other situations the independent white noise assumption has proven quite realistic.

Let us consider the case of a hidden mosaic process \mathbf{x} with independent Gaussian observations \mathbf{y} on top of it, say $y_i | \{x_i = k\} \sim f(y_i | x_i) = \mathcal{N}_d\{\mu(k), \Sigma\}$ for definiteness. The restoration problem is to estimate the full image \mathbf{x} from the observed \mathbf{y} . This can be viewed as a spatial classification task. The simplest solution is to carry out ordinary discriminant analysis for each pixel, that is, use as \hat{x}_i the class label k that maximises $\pi(k)f(y_i | k)$, where $\pi(1), \dots, \pi(K)$ are the prior probabilities for the K classes. Of course the parameters $\mu(1), \dots, \mu(K)$, Σ must have been estimated from some initial training stage. In some applications the class densities have large pairwise interdistances and this simple non-contextual method is sufficient. In other applications it is not, and contextual methods can offer significant improvements. We discuss two major approaches below, the “local modelling”

developed by Hjort & Mohn (1984, 1987) and others and the mrf based one developed by Geman & Geman (1984), Besag (1986) and others. Other approaches are mentioned in section 3.4.4.

3.4.2. An approach based on neighbourhood models

Let us see where basic statistical decision theory leads us in the present spatial context. Assume that the loss incurred when we assign label \hat{x}_i to pixel i , whose true label is x_i , is of the type

$$L(x_i, \hat{x}_i) = \begin{cases} 0 & \text{if } \hat{x}_i = x_i \text{ (correct decision),} \\ 1 & \text{if } \hat{x}_i \neq x_i \text{ and } \hat{x}_i \in \{1, \dots, K\} \text{ (wrong decision),} \\ t & \text{if } \hat{x}_i = D \text{ (being in doubt),} \end{cases} \quad (3.21)$$

in which t is a threshold between 0 and 1. Thus the possibility of being in doubt and state nothing about x_i is reserved, having in mind, for example, mixed pixels in some remote sensing application. Now, if the total measure of consequence is the average of the individual loss-contributions, i.e. the misclassification rate plus t times the doubt rate, then the optimal rule becomes

$$\hat{x}_i = \begin{cases} k & \text{if } k \text{ maximises } P_i\{k | y\}, \text{ and this maximum exceeds } 1 - t; \\ D & \text{if } P_i\{k | y\} \leq 1 - t \text{ for each } k. \end{cases} \quad (3.22)$$

Two points to note here are that all the data in principle are conditioned on in the posterior probabilities $P_i\{k | y\}$, and that the rule classifies one pixel at a time.

In practice one has to limit oneself to a small subset $y_{N(i)}$, containing at least y_i , of all the data. If $y_{N(i)}$ is chosen to consist of y_i and its four immediate neighbours, for example, then the natural approximation to (3.22) is the rule that maximises

$$\begin{aligned} P_i\{k | y_{N(i)}\} &= \text{const } \pi(k) f(y_{N(i)} | c_i = k) \\ &= \text{const } \pi(k) \sum_{a, b, c, d} g(a, b, c, d | k) h(y_{N(i)} | k, a, b, c, d). \end{aligned} \quad (3.23)$$

Here the two basic stochastic elements of the problem enter in a natural and illuminating way: $g(a, b, c, d | k)$ is the conditional probability of getting class configurations a, b, c, d given class k in the centre pixel, and is tied to the \mathbf{x} process. And $h(y_{N(i)} | k, a, b, c, d)$ is the simultaneous density for the five vectors in question, given that the underlying classes are k, a, b, c, d . The summation in (3.23) is, in general, over all K^4 configurations. The rule based on (3.23) is an approximation to (3.22), but we emphasise that it also enjoys a natural optimality property by itself, namely that of achieving lowest expected average loss among all rules using the neighbour information $y_{N(i)}$ for the i th decision.

For each specification of global, simultaneous models for the processes \mathbf{x} and for \mathbf{y} given \mathbf{x} formulae for g and h above can in principle be derived, after which we have a contextual classification algorithm. It is not really necessary for us to derive g and h from fully given, simultaneous probability distributions, however; we may if we wish forget the full scene and come up with realistic local models for the pixel neighbourhood alone, i.e. model g and h above directly. Even if some proposed local g -model should turn out to be inconsistent with a full model for \mathbf{x} , say, we are allowed to view it merely as a convenient approximation to the complex schemes Nature employs when she distributes class labels over the scene.

Another typical feature in these problems is also illustrated in formula (3.23): the models we use must not only be realistic and fitable, but also feasible in the sense of not needing too much computing time. A satellite scene can contain about a million pixels, and a rule that needs to sum K^4 terms for each class before it can decide on a class label for a pixel will be useless in most cases. Accordingly we should look for clever approximations and/or for convenient model choices that lead to reduced and simplified expressions. One such clever version of the general (3.23) is the following: suppose that y_i -vectors given the underlying classes are independent, and let $g(a, b, c, d | k) = p(a | k)p(b | k)p(c | k)p(d | k)$, where the $p(a | k)$ s are neighbour transition probabilities (which must be estimated). The Markov mesh model for classes on a lattice studied by Pickard (1977) has in fact this multiplicative property. Then (3.23) simplifies to

$$P_i\{k | y_{N(i)}\} = \text{const } \pi(k)f(y_i | k)T_k(y_{i1}) \cdots T_k(y_{i4}), \quad T_k(y) = \sum_{m=1}^K p(m | k)f(y | m), \quad (3.24)$$

where the y_{ij} s stem from the four neighbour pixels. This produces the classification rule reached by Hjort & Mohn (1984), Haslett (1985), and others, from somewhat different perspectives.

Hjort & Mohn (1985) obtained a natural generalisation of this rule to the case where spatial autocorrelation between y_i s is allowed for. Specifically,

$$P_i\{k | y_{N(i)}\} = \text{const } \pi(k)f(y_i | k)U_k(y_i, y_{i1}) \cdots U_k(y_i, y_{i4}), \quad (3.25)$$

where the U -functions are appropriate generalisations of the T -functions appearing in (3.24). They also provide evidence that such autocorrelation is prominently present with high resolution satellite data and ought to be taken into account. Hjort & Mohn (1984) and Sæbø *et al.* (1985) consider other variations on theme (3.23) as well.

The reasoning above applies equally well to larger neighbourhoods than the cross, but exact expressions based on the appropriate generalisation of (3.23) quickly become long and untractable. Again we feel that the statistician should not be afraid of constructing pragmatic approximations, even if they should lead him outside the safe ground of exact expressions under exact models. For the 3×3 pixel box with eight neighbours, for example, we may use a formula similar to (3.24), with four more terms entering the product, involving transition probabilities for diagonal neighbours, say $q(\cdot | k)$, which can be expressed in terms of the one-step neighbour transition probabilities $p(\cdot | \cdot)$ s. This produces a valid classification algorithm with good error rate properties, although it, in fact, cannot be deduced from a bona fide global model for the classes.

A natural question is how much gain there is in being (more) sophisticated and cpu-consuming, by including neighbours at all, and by including say eight neighbours instead of only four. Hjort (1985a) studies one particular eight-neighbour method of the type (3.23), by an appropriate generalisation of a four-neighbour method due to Owen (1984). Exact formulae for error rates cannot be obtained, but they can be expressed via probabilities for events involving nine or fewer independent normal variates (univariate, even if the y_i s are multivariate), and can as such be evaluated by computer simulation. (In this way we do not have to simulate the scenes or portions of the scene itself.) Some numerical information is presented in Hjort & Mohn (1987), and indicates first of all that using context can lead to appreciably increased accuracy, and secondly that using larger neighbourhoods usually will be worth the extra trouble and cpu-time. This is also supported by experience from simulation studies, see for example Hjort *et al.* (1987).

3.4.3. Markov random field approaches

The loss function (3.21) is local in nature, and corresponds essentially to viewing pixel-wise error rate as the basic quality measure. A radically different suggestion is the global loss measure

$$L(\mathbf{x}, \hat{\mathbf{x}}) = \begin{cases} 0 & \text{if every pixel is correctly classified;} \\ 1 & \text{if one or more pixels are misclassified.} \end{cases} \quad (3.26)$$

The optimal rule in this case becomes: find $\hat{\mathbf{x}} = (\hat{x}_1, \dots, \hat{x}_N)$ to maximise the posterior probability

$$p(\mathbf{x} | \mathbf{y}) = \text{const } p(\mathbf{x})f(\mathbf{y} | \mathbf{x}).$$

This requires first of all full model specifications for \mathbf{x} and \mathbf{y} (as opposed to only “local models”). Secondly, it may appear practically impossible to find this maximum *a posteriori* probability scene, because of the enormous number K^N of different possibilities to search through. But modern ideas from numerical–probabilistic optimisation made it possible for both Geman & Geman (1984) and Besag (1986), in two important papers, to give satisfactory solutions, for the broad family of mrf prior distributions studied in section 3.2.

Consider for illustration the simplest type of mrf (3.17), with a single β parameter in addition to class-parameters $\alpha(k)$. Assume also that the y_i s are conditionally independent. Then

$$\begin{aligned} p(\mathbf{x} | \mathbf{y}) &= \text{const } p(\mathbf{x})f(\mathbf{y} | \mathbf{x}) \\ &= \text{const } \exp \left[\sum_{i=1}^N \{\alpha(x_i) + \log f(y_i | x_i)\} + \beta \sum_{i < j} I\{x_i = x_j\} \right]. \end{aligned}$$

Accordingly \mathbf{x} given data is again a mrf, only with updated $\alpha(x_i)$ s. Geman & Geman (1984) discuss a “statistical cooling” technique from combinatorial optimisation for coming at least close to the maximum *a posteriori* picture, which is the optimal solution w.r.t. loss function (3.26). It is computationally demanding and requires several hundred iterative scans over the full scene. See Marroquin *et al.* (1987) for considerations about massive parallel processors and speed, and for applications of mrfs to computer vision. Besag (1986), on the other hand, has proposed a much simpler computational scheme that in effect, in a coordinate-ascent way, goes for a local maximum of the posterior distribution. The intuitively plausible idea is to start with an initial estimate $\bar{\mathbf{x}}$ for the scene, and then update $\bar{\mathbf{x}}$ to a perhaps better $\hat{\mathbf{x}}$ by finding $\hat{x}_i = k$ to maximise

$$p_i(x_i = k | \mathbf{y}, \hat{\mathbf{x}}_{S-i}) = \text{const } \exp [\alpha(k) + \beta H_i(k, \bar{x}_{\delta i})] f(y_i | k), \quad (3.27)$$

with notation as in (3.17). Note that this is like ordinary discriminant analysis but with neighbour-influenced “prior” probabilities $\pi(k) = \text{const } \exp [\alpha(k) + \beta H_i(k, \bar{x}_{\delta i})]$. In this way the full scene is swept over, in some order, and we have a new, updated estimate $\hat{\mathbf{x}}$. The process is iterated until convergence; usually 6–10 times suffice. The starting point is ordinarily that corresponding to $\beta = 0$, i.e. the non-contextual, but the “iterated conditional modes” method could equally well use a contextual classification as its starting point. It has been argued that smaller values of β should be used for the first couple of iterations.

It is our experience from high resolution satellite data that realistic models must allow positive spatial correlation for y_i -vectors given the scene. A simple model that serves to illustrate the general principle and the wider potential of the mrf approach is the following:

$$f(\mathbf{y} \mid \mathbf{x}) = \text{const exp} \left\{ -\frac{1}{2} \sum_i (y_i - \mu(x_i))' \Sigma^{-1} (y_i - \mu(x_i)) \right. \\ \left. - \sum_{i < j} \gamma_{ij} (y_i - \mu(x_i))' \Sigma^{-1} (y_j - \mu(x_j)) \right\},$$

in which $\gamma_{ij} = \frac{1}{2}\gamma$ when i and j are immediate (first order) neighbours and zero otherwise. This is a Gaussian mrf, or conditional autoregressive scheme, with corresponding local characteristics

$$y_i \mid \mathbf{x}, \mathbf{y}_{S-i} \sim N\{\mu(x_i) + \gamma(\bar{y}_{\partial i} - \bar{\mu}_{\partial i}), \Sigma\}, \quad (3.28)$$

writing $\bar{y}_{\partial i}$ for the average of the four immediate y_j neighbours to y_i and similarly $\bar{\mu}_{\partial i}$ for the average of the four accompanying $\mu(x_j)$ s. It is easily seen that \mathbf{x} given \mathbf{y} again is a mrf. An appropriately modified version of the Geman and Geman method is capable of coming close to the simultaneous optimisation of $p(\mathbf{x} \mid \mathbf{y})$, which is seen to mean maximising

$$\sum_i [\alpha(x_i) - \frac{1}{2}(y_i - \mu(x_i))' \Sigma^{-1} (y_i - \mu(x_i))] \\ + \sum_{i < j} [\beta I\{x_i = x_j\} + \frac{1}{2}\gamma(y_i - \mu(x_i))' \Sigma^{-1} (y_j - \mu(x_j))],$$

where the second sum is over all pairs of neighbours. The natural generalisation of Besag's method to the present spatial correlation model is simpler, and amounts to maximising for each i , and for the current estimate $\bar{\mathbf{x}}$ of the rest of the scene,

$$p_i(x_i = k \mid \mathbf{y}, \bar{\mathbf{x}}_{S-i}) = \text{const exp} [\alpha(k) + \beta H_i(k, \bar{x}_{\partial i})] f(y_i \mid k, \bar{\mathbf{x}}_{S-i}, \mathbf{y}_{S-i}). \quad (3.29)$$

Thus the conditional mode step again acts like discriminant analysis, but this time with both neighbour-corrected prior probability and neighbour-corrected class density.

3.4.4. Other approaches to classification

Let us briefly touch upon some other spatial classification techniques. Underlying the previous methods are the loss functions (3.21) and (3.26). A simple intermediate loss function that is less crude than (3.21) but nevertheless has a "contextual element" is $L = \text{number of misclassified } 2 \times 2 \text{ blocks of pixels}$. The optimal rule becomes one of simultaneously classifying four pixels, by maximising the posterior probability given data over the K^4 possible outcomes. Hjort (1987) has worked out simple rules of this type based on a geometric probability model for classes, under which only $6K^2 - 5K$ of the possible configurations have positive probability. This geometrical model used the Switzer process for Poisson lines mentioned in section 3.2.5.

Relaxation procedures is a term used for algorithms that iteratively adjust posterior probabilities based on estimated or known spatial relationships among the class labels. Some of these use wider and wider neighbourhoods as the iterations go on. They are perhaps best understood in terms of connections to mrf models, and to analysis of incomplete data, see Kay & Titterington (1986) and Fiskum (1986), and section 4.2 below.

Other references of interest include Switzer & Green (1984) and Switzer & Ingebritsen (1986) on min/max autocorrelation factors that aim to separate noise from signal; Conradsen & Nielsen (1987), studying the benefits of using texture-type features derived from neighbourhoods; Green & Titterington (1987) who study recursive procedures under an interesting model for sequences of images; Esbensen & Geladi (1989) where soft bi-linear modelling is used; Greig *et al.* (1989) who obtain and demonstrate an exact maximum *a posteriori* image algorithm for the two class case; Owen (1989), where the smoothing parameters of Besag

type methods are studied; and Taxt & Bølviken (1991), where new restoration algorithms are motivated through an analogy with quantum physics.

3.4.5. Predicting a continuous variable

In several remote sensing applications one is as interested in estimating a ground parameter as in classifying pixels into ground classes. Imaging that z_i is such a variable of interest, associated with ground surface element i . In a water quality application of MSS- and LANDSAT-satellite data we have worked on there are in fact several z_i 's of interest: the amount of plankton in sea element i , along with turbidity, water transparency, and other measures of water quality. In an application to forestry surveillance the total tree mass volume for each 20 m × 20 m element on the ground was of interest. With luck these z_i 's are correlated with the remotely sensed y_i -vectors, and predictions based on these can be made. Examples of this sort abound in the remote sensing literature, and make it clear that many important surveillance tasks can now be carried out with the help of remote sensing, in proper combination with ground truth measurements from land stations.

Let us for convenience still use y_i to denote the vector of pertinent observations at pixel i , for example spectral data, possibly transformed, and possibly supplemented with other available covariates thought to be useful for the prediction of z_i . (In the water quality application, the component of y_i are two chromaticity indices supplemented with topographic information.) Let us also keep z_i one-dimensional (extensions are straightforward). A useful mathematical assumption, used as a vehicle for producing good prediction procedures, is

$$\begin{pmatrix} z_i \\ y_i \end{pmatrix} \mid (x_i = k) \sim \mathcal{N}_{d+1} \left\{ \begin{pmatrix} v_k \\ \mu_k \end{pmatrix}, \begin{pmatrix} \tau_k^2 & \omega'_k \\ \omega_k & \Sigma \end{pmatrix} \right\}.$$

Thus z_i in pixel i is viewed as a realisation of a random variable. The natural predictor, if we know that the pixel in question is of ground type k , is

$$\bar{z}_i(k) = E\{z_i \mid y_i, x_i = k\} = v_k + \omega'_k \Sigma^{-1}(y_i - \mu_k), \quad (3.30)$$

and the conditional variance $\sigma(k)^2 = \tau_k^2 - \omega'_k \Sigma^{-1} \omega_k$ can be used to construct prediction intervals for z_i . Without knowledge of the pixel's class one might weight these with the posterior probabilities $P_i\{k \mid y_i\}$ to form a (non-contextual) predictor.

There are several ways to incorporate spatial context in the method. One strategy that uses a spatial model for a continuous residual function is explained and illustrated in Høst *et al.* (1989). Let us now examine some other contextual alternatives that use spatial models for the underlying classes.

3.4.6. Predictors constructed from neighbourhood models

Hjort & Mohn (1987) derived prediction rules with variance measures for this framework. Assume that z_i 's given y and x are independent, with $z_i \mid (y_i, x_i) \sim f(z_i \mid y_i, x_i)$, a normal with mean $\hat{z}_i(x_i)$ and variance $\sigma(x_i)^2$. The predictors take the form

$$\hat{z}_i = E\{z_i \mid y_{N(i)}\} = \sum_{k=1}^K P_i\{k \mid y_{N(i)}\} \hat{z}_i(k).$$

One readily shows that z_i given data $y_{N(i)}$, i.e. the information from pixel i and its neighbours, is distributed as a mixture of the densities $f(z_i \mid y_i, x_i)$ with $P_i\{k \mid y_{N(i)}\}$ as weights. Accordingly

$$\text{var}\{z_i \mid y_{N(i)}\} = \sum_{k=1}^K P_i\{k \mid y_{N(i)}\} \{\sigma(k)^2 + (\hat{z}_i(k) - \hat{z}_i)^2\},$$

which is of use when confidence intervals are called for. Observe that if a contextual classification is carried out using the local modelling method then the extra computational burden needed to compute predictors and standard deviations is mild.

3.4.7. Markov random field methods

Assume that \mathbf{x} is a mrf with distribution (3.14), that \mathbf{y} given \mathbf{x} is a Gaussian mrf with local characteristics (3.28), and that z_i s are conditionally independent as above. Then one can show that (\mathbf{y}, \mathbf{z}) has a simultaneous mrf distribution given \mathbf{x} , and that $(\mathbf{x}, \mathbf{y}, \mathbf{z})$ is a simultaneous mrf. The most important fact is that the unobserved processes (\mathbf{x}, \mathbf{z}) given the data \mathbf{y} , the Bayesian posterior distribution, is yet another mrf, with local characteristics

$$(x_i, z_i) | (\mathbf{y}, \mathbf{x}_{S-i}, \mathbf{z}_{S-i}) \sim \text{const} \exp [\alpha(x_i) + \beta H_i(x_i, x_{\partial i})] f(y_i | \mathbf{x}, \mathbf{y}_{S-i}) f(z_i | y_i, x_i), \quad (3.31)$$

cf. (3.17) and (3.28). This generalises the result (3.29).

Besag's method of taking iterative conditional modes can be generalised to the present state of affairs, with both \mathbf{x} and \mathbf{z} to be estimated from the image data \mathbf{y} . This is done in Hjort & Mohn (1987, sect. 4). The best method depends on the specific loss function used. When the loss function is $I\{\hat{\mathbf{x}} \neq \mathbf{x}\}$ plus average squared prediction error then the Bayes solution is as follows: first use the usual Besag method (in this case, as explained around (3.29)) to arrive at the scene estimate $\hat{\mathbf{x}}$. Then compute

$$\hat{z}_i = E\{z_i | \mathbf{y}, \hat{\mathbf{x}}_{S-i}, \mathbf{z}_{S-i}\} = \frac{\sum_{k=1}^K \exp [\alpha(k) + \beta H_i(k, \hat{x}_{\partial i})] f(y_i | k, \hat{\mathbf{x}}_{S-i}, \mathbf{y}_{S-i}) \hat{z}_i(k)}{\sum_{k=1}^K \exp [\alpha(k) + \beta H_i(k, \hat{x}_{\partial i})] f(y_i | k, \hat{\mathbf{x}}_{S-i}, \mathbf{y}_{S-i})}.$$

Note that the autocorrelation parameter γ enters via $f(y_i | \mathbf{x}, \mathbf{y}_{S-i})$.

In this case \hat{z}_i emerged as an explicit function of the finally classified scene $\hat{\mathbf{x}}$. This is because we assumed conditional independence of z_i s given \mathbf{x} and \mathbf{y} . In models where z_i s must be taken interdependent, given scene and \mathbf{y} -data, a simultaneous, iterative updating of \mathbf{x} and \mathbf{z} may be called for. An example is given in Hjort & Mohn (1987, sect. 4).

3.4.8. Marked point processes

Suppose there is a mark or set of attributes z_i associated with each point x_i of a spatial point process. An example with a 4-dimensional mark for each point is described in section 4.3 below. If the marks live in a mark space \mathcal{Z} then the process with outcomes (x_i, z_i) is just a spatial point process in some appropriate $\mathcal{X} \times \mathcal{Z}$, so the most important parts of the theory presented in section 3.3 carry over to marked point processes. In particular, simulation of realisations can be carried out using a spatial birth and death process. The hardest and most vital task is often simply that of building a good model, with distance functions and pairwise interaction functions, that produces realistic outcomes. A general reference with theory and example is Stoyan *et al.* (1987).

3.4.9. Estimation problems

There are challenging estimation problems associated with several of the models described. Hjort (1985b) and Hjort & Mohn (1987) develop estimation methods for many of the image models mentioned in section 3.4.2. In particular they describe methods that utilise unclassified

vectors via estimation of mixture distributions. Besag (1986) proposes an iteration scheme to restore an image and simultaneously estimate the parameters of the mrf model used for \mathbf{x} , used in section 3.4.3, assuming the y_i s to be conditionally independent. This scheme is somewhat biased and inconsistent, as pointed out along with a remedy in Hjort & Mohn (1987). Lakshmanan & Derin (1989) describe a simulated annealing method that stops at regular intervals to estimate the mrf parameters. Veijanen (1990) gives another method for imperfectly observed mrfs and proves consistency. A deep treatment for maximum pseudo-likelihood estimation with hidden mrfs is given by Comets & Gidas (1992). Georgsen & Omre (1993) consider estimation in a model that combines fibre processes with a Gaussian random function.

4. Some worked-through examples of applications

4.1. Depth conversion of seismic data (Omre & Halvorsen, 1989; see also Abrahamsen *et al.*, 1991 and Abrahamsen, 1993)

The petroleum reservoirs in the North Sea are located at a depth of approximately 3000 m and have an areal extent of typically $3.0 \times 5.0 \text{ km}^2$. In the reservoir the hydrocarbons tend to migrate upwards in the structures until they are trapped under a syncline of non-permeable geologic layers, usually shale-rich horizons. The mapping of these horizons is important both for exploration purposes and for prediction of hydrocarbon volume *in situ*.

Fortunately these non-permeable horizons tend to have properties very different from the porous reservoir, and can be identified from seismic data. Seismic data can be collected in an inexpensive way from ships, and vast amounts with good areal coverage are usually available. The seismic reflection signal has geographical reference horizontally and reflection time reference vertically. After seismic cleaning, which is a discipline in itself, a relatively reliable seismic reflection time surface $\{t(x): x \in D\}$ is obtained, see Fig. 1. The challenge is to convert this into a map of depths to the horizon $\{z(x): x \in D\}$. This could of course be done by simply multiplying reflection time and signal velocity. The fact that the signal velocity varies considerably, both vertically and laterally, complicates this.

Two other sources of information are available for the depth conversion. The first is the knowledge of the geophysicists. The areal extent of some of these horizons is often huge and covers other reservoirs. Hence experience from other areas of the North Sea is relevant. In addition the physical understanding of packing of reservoir rock provides some constraints on the vertical velocity profile. Secondly, exact depth observations $z(x_1), \dots, z(x_n)$ are available in certain well positions, see Fig. 1. The fact that the four first wells are the most shallow ones indicates preferential drilling and non-representative positioning of the observations in the design space. Note, however, that the seismic reflections have a support of $100 \times 100 \text{ m}^2$ of the horizon while the support of well observations is $0.2 \times 0.2 \text{ m}^2$. Hence the former must be considered as weakly smoothed.

The example is from a study of a huge offshore gas field in the North Sea. A set of parameters, intended to be as realistic as possible, were defined in cooperation with geophysicists in a Norwegian oil company. The sensitivity to the influence of the prior knowledge was evaluated. The changes in the predictions over time, i.e. as a function of the available wells, were also studied. The stochastic model used for merging the various pieces of information in this study is of the form

$$Z(x) = V(x)t(x) + \varepsilon(x) = \left(\beta_1 + \beta_2 t(x) + \beta_3 \frac{x_1 - x_{1,\min}}{x_{1,\max} - x_{1,\min}} \right) t(x) + \varepsilon(x), \quad (4.1)$$

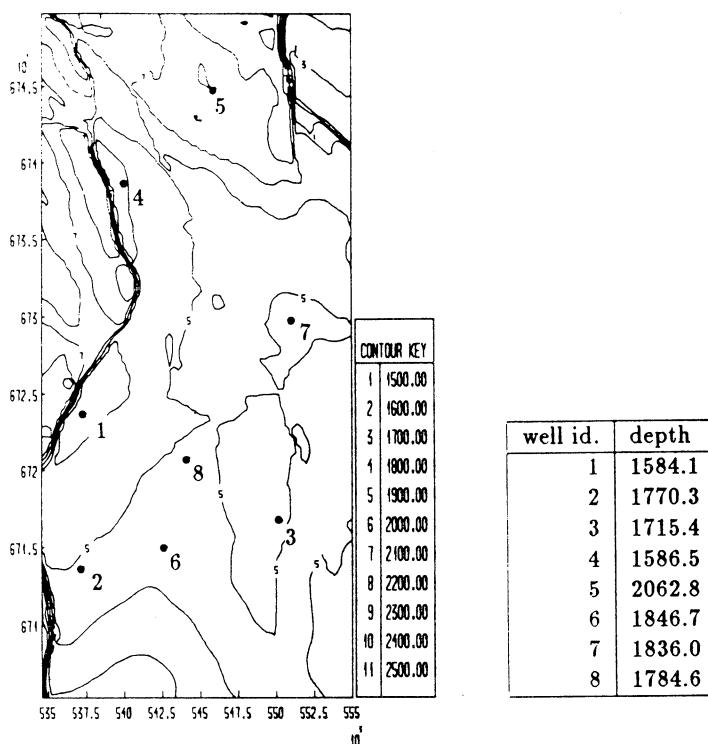


Fig. 1. Map of seismic reflection times and list of depths observed in the wells.

where $\varepsilon(x)$ is a continuous residual surface, compensating for the smoothing in the seismic signal, with correlation structure $\sigma^2 K(\|x - y\|)$. In other words, the signal velocity increases with depth according to the increase in reflection time and there exists a lateral trend in the West–East or x_1 -direction. Geophysical knowledge is included through carefully assessed prior distributions on $\beta_1, \beta_2, \beta_3$. The model is accordingly within the Bayesian Kriging framework as discussed in section 3.1.3. The spatial covariance function used is

$$K(r) = \begin{cases} 1 - \frac{3}{2} \frac{r}{2800} + \frac{1}{2} \left(\frac{r}{2800} \right)^3 & \text{if } 0 \leq r \leq 2800, \\ 0 & \text{if } r \geq 2800, \end{cases}$$

where r is the distance from x to y , and the σ parameter was estimated to be 200.0 m. This estimate was based on estimates in comparable reservoirs with numerous wells and on understanding of the physics of seismic reflections. The prior distribution was assessed in cooperation with experienced geophysicists, and is as follows:

$$\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} \sim \mathcal{N}_3 \left\{ \begin{bmatrix} 0.600 \\ 0.0002 \\ 0.070 \end{bmatrix}, \tau^2 \begin{bmatrix} (0.055)^2 & 0 & 0 \\ 0 & (0.0001)^2 & 0 \\ 0 & 0 & (0.022)^2 \end{bmatrix} \right\}. \quad (4.2)$$

The 0.055, 0.0001, 0.022 values were the actual prior standard deviation parameters used, but the scale parameter τ was also included and varied in order to evaluate the sensitivity to the choice of prior distribution; see also the relevant discussion in Omre & Halvorsen (1989).

For τ equal to zero the coefficients β are fully specified and the underlying trend surface can be subtracted. This corresponds to what is termed simple Kriging in geostatistical

Table 1. Posterior distribution of $(\beta_1, \beta_2, \beta_3)$, as a function of the number of wells used in the conditioning. The table gives posterior means, standard deviations, and correlation coefficients. The correlations are respectively between β_1 and β_2 , between β_1 and β_3 , and between β_2 and β_3 .

No. of wells	Means	SD	Correlations
0	0.6000	0.0548	0
	0.00020	$100.0 \cdot 10^{-6}$	0
	0.0700	0.0223	0
1	0.5942	0.0523	-0.991
	0.00017	$29.9 \cdot 10^{-6}$	0.101
	0.0699	0.0221	-0.223
3	0.5865	0.0491	-0.990
	0.00018	$26.7 \cdot 10^{-6}$	-0.051
	0.0501	0.0123	-0.034
5	0.5581	0.0362	-0.988
	0.00018	$19.6 \cdot 10^{-6}$	0.157
	0.0532	0.0119	0.290
8	0.5654	0.0356	-0.991
	0.00018	$19.1 \cdot 10^{-6}$	0.101
	0.0663	0.0101	-0.221

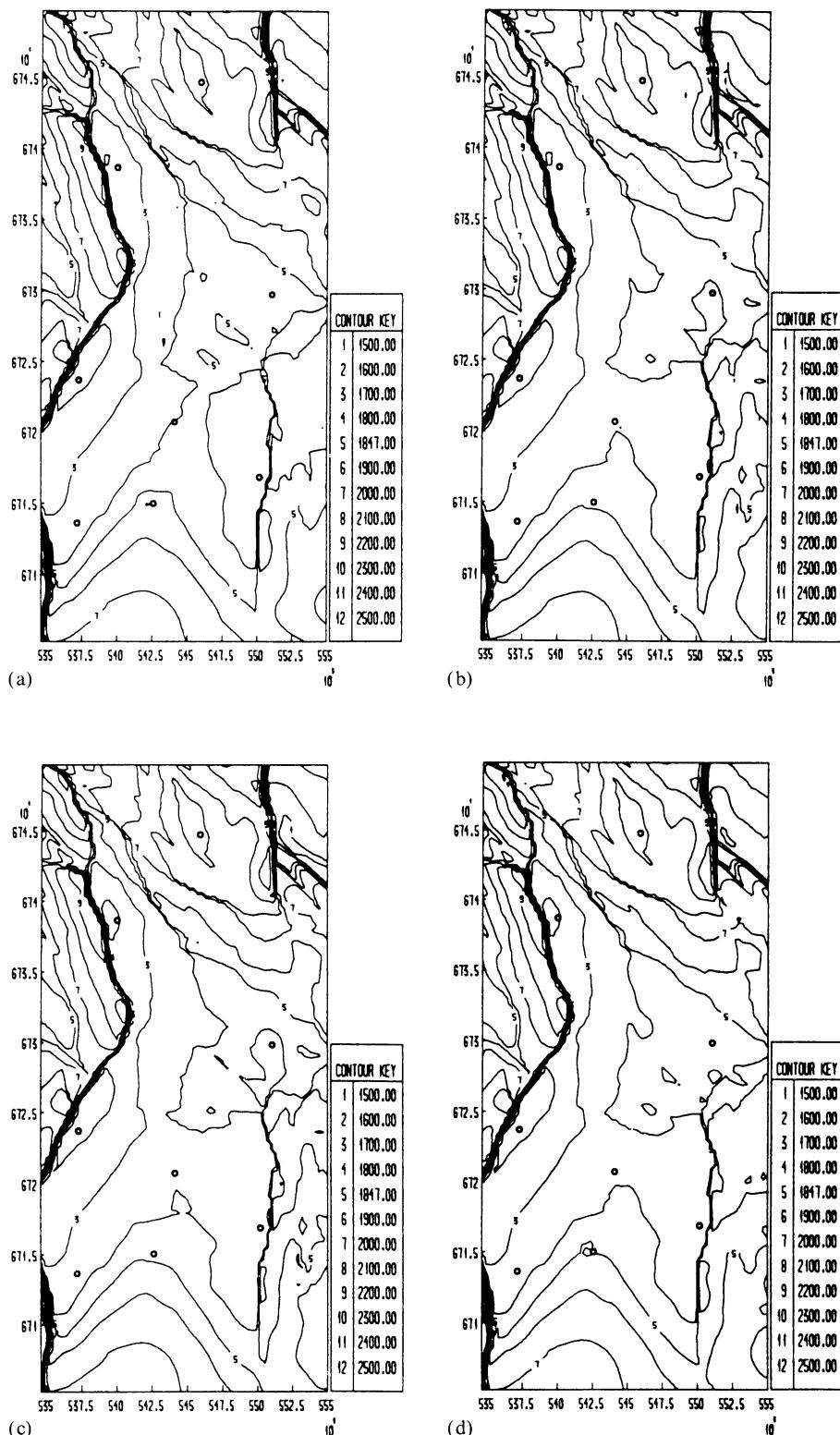
terminology. The observations $z(x_1), \dots, z(x_n)$ will in this case have influence on the residuals $\{e(x) : x \in D\}$ only. If the parameters are misspecified this makes the observations create cone-like surfaces in their immediate neighbourhood. The case of τ going to infinity corresponds to there being no available prior knowledge about the coefficients, and the resulting method coincides with what is called universal Kriging. In this case the observations $z(x_1), \dots, z(x_n)$ will be the only source of information for both $\beta_1, \beta_2, \beta_3$ and the residuals. This may give unreliable results based on the first few wells since they are known *a priori* to be preferentially located in shallow areas. The intermediate cases with finite τ constitute trade-off cases between user experience and the available observations, as discussed in section 3.1.

The development of the predictions over time can be evaluated by adding one well at a time. In Table 1 and Fig. 2 results from this study are summarised. The prior distribution actually used is the one with $\tau = 1$.

Note that the model in this application is used for interpolation. When assessing the hydrocarbon volume *in situ* with associated uncertainty measures one would use the model as basis for stochastic simulation.

The regression surface contains three unknown parameters, so a solution in the traditional setting requires at least three wells. In the Bayesian setting a solution exists regardless of the number of observations. The Bayesian model provides a prior guess on the surface, along with uncertainty measures, and the initial estimate is updated when new observations are made available. From an operational point of view in petroleum exploration this is meaningful, and it provides the opportunity to evaluate the information content in each datum. Note also that the posterior variances of the parameters decline monotonically with increasing number of wells, as they should, and that in the posterior distribution the off-diagonal terms in the covariance matrix for β are non-zero, since all the parameters draw on the information in the observed wells.

We view the use of the Bayesian machine here as appropriate and non-controversial, in spite of the somewhat subjective and surely non-perfect determination of the prior distribution. The geophysical knowledge is substantial and is of a form which naturally can be quantified through the parameters of the model as defined above. A lot of expert effort and



experience has gone into the determination of the prior parameters. The setting of these was not an entirely Bayesian affair but had more of the empirical Bayes flavour, in that also other available indirect data sources were used, partly in informal ways via plots etc.

Reliable estimators for the spatial correlation function $K(\cdot)$ and the standard deviation parameter σ are not simple to obtain with so few data points available. Often experience from evaluation of other reservoirs with many more wells must be exploited. This suggests using the extended Bayesian apparatus of the type described in 3.1.3, with an inverse gamma prior distribution for σ^2 and a “hyper prior” distribution for τ^2 , and perhaps with a prior for the covariance matrix of β . In Abrahamsen *et al.* (1991) and Abrahamsen (1993) the model is extended to solve a multi-layer seismic depth inversion problem, providing solutions for observations from non-vertical wells.

4.2. Identification of heart dysfunction (Taxt *et al.*, 1990)

The objective of the study is to evaluate the left ventricular volume of the heart and the cardiac output fraction based on data from time-repeated two-dimensional echocardiography, so-called ultrasound techniques. This is expected to lead to significant improvements of present diagnostic capabilities for heart dysfunction. The methodological challenge is to combine prior knowledge with pattern recognition techniques to identify moving boundaries of specific structures in time-varying images with low local contrast and a large noise component.

Time-varying ultrasound images are extensively used as a routine diagnostic tool in the examination of the cardiac function. The images are collected at 25–35 Hz which is appropriate for evaluating the heart activity with contraction cycles down to 0.3 s. Each image is collected in polar coordinates (r, θ) , with the origin at the probe of the ultrasound equipment. The reflections u are assigned integer-values in the range from 0 to 255. In the study 20 consecutive images were available, covering more than two complete cardiac cycles. The data are accordingly of the form $u_t(r_i, \theta_j) \in \{0, \dots, 255\}$ for $i = 1, \dots, 512$, $j = 1, \dots, 128$, $t = 1, \dots, 20$. Since the reference system is not orthogonal an inverse polar transformation was used to get $u_t(x_i, y_j)$ instead. The time reference was treated as a third dimension in the problem, hence defining a 3-D image. To reduce the computational load the grey level resolution was reduced to 32. Thus data are of the form

$$v_{ijk} = u_k(x_i, y_j) \in \{0, \dots, 31\} \quad \text{for } i = 1, \dots, N_i, j = 1, \dots, N_j, k = 1, \dots, 20. \quad (4.3)$$

In Fig. 3 a cross section of the 3-D image at constant k is displayed, i.e. at a specific time point.

Two noise reduction algorithms have been used in this study. A relaxation procedure of the type mentioned in section 3.4 is the basis for the first one. The relaxation procedure is adapted to the noise reduction problem as follows: use as (one-dimensional) feature vector the grey level itself; use as the set of possible classification classes simply the possible grey levels $c \in \{0, \dots, 31\}$; and use class-conditional feature vector density $f_c \sim \mathcal{N}\{c, \sigma^2\}$ combined with transition probabilities $p(d | c) = \frac{1}{31}\{1 - p(c | c)\}$ for $d \neq c$.

Fig. 2. Bayesian Kriging maps, conditioning on respectively (a) one well, (b) three wells, (c) five wells, and (d) eight wells.

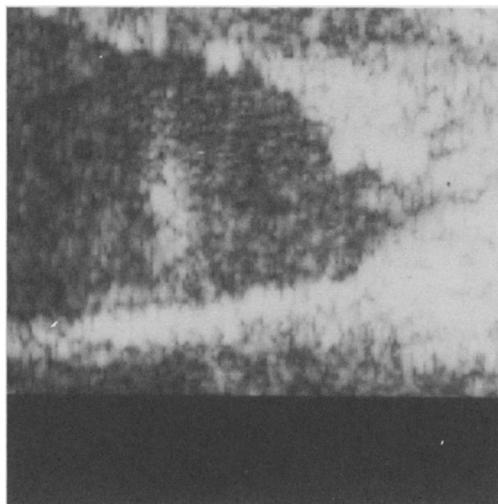


Fig. 3. Cross section at a given point in time of the cardiac output of a part of the heart in true coordinates.

To describe how the relaxation simulation scheme works, let C'_{ijk} denote temporary classification for pixel (i, j, k) at iteration l , let V_{ijk} be a neighbourhood of (i, j, k) , and let V^2_{ijk} be the neighbours plus all their neighbours again. Then

$$\begin{aligned} \Pr \{C'_{ijk}^{l+1} = c \mid \text{all } C'_{mno} \text{ for } (m, n, o) \in V^2_{ijk}\} &= \text{const exp}\{-(c - C'_{ijk})^2/2\sigma^2\} \\ &\times \sum_{(m, n, o) \in V_{ijk}} \left[p(c \mid c) \Pr \{C'_{mno} = c \mid \cdot\} + \frac{1}{31} \{1 - p(c \mid c)\} [1 - \Pr \{C'_{mno} = c \mid \cdot\}] \right]. \end{aligned} \quad (4.4)$$

The initiation of the procedure is done by determining $\Pr \{C^0_{ijk} = c \mid \cdot\}$ from a non-contextual Gaussian model assuming $\pi_c = 1/32$, and using of course $c^0_{ijk} = v_{ijk}$ in all positions as the initial image. The parameter values used in the actual study were $p(c \mid c) = 0.90$ and $\sigma = 7.41$. The former was subjectively assigned while the latter was estimated from a training set of data. The neighbourhood V_{ijk} used is of size $3 \times 3 \times 3$.

The relaxation was carried out in five iterations, after which each pixel was assigned the most probable grey level according to the estimated posterior probabilities. The resulting noise reduced image for a particular time-index is displayed in Fig. 4. This image is much smoother than the original one in the sense that noise seems to have been removed. Note that the edges seem to be partially smoothed away as well, however.

Besag's method of iterated conditional modes, as presented in section 3.4, provides the basis for the other noise reduction algorithm. Corresponding adaptations as for the relaxation procedure were made. The archetypal Besag model contains class-parameters $\alpha_1, \dots, \alpha_{32}$ and a single β for neighbourhood interaction. The α_c s were taken class independent in the study reported on here. The resulting Besag iteration equation becomes

$$\Pr \{C'_{ijk}^{l+1} = c \mid C'_{mno} \text{ for } (m, n, o) \in V_{ijk}\} = \text{const exp} \left\{ \frac{-(c - C'_{ijk})^2}{2\sigma^2 + \beta H'(c, V_{ijk})} \right\}, \quad (4.5)$$

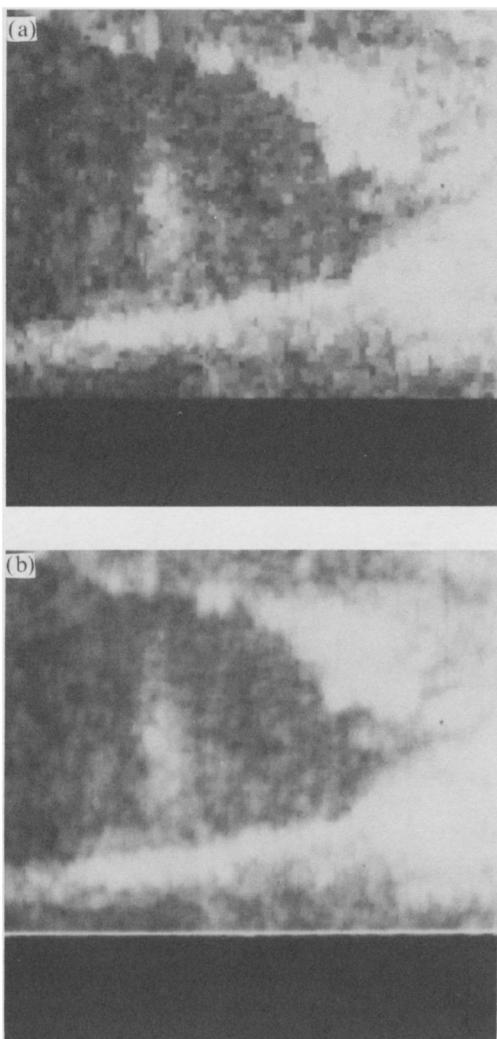
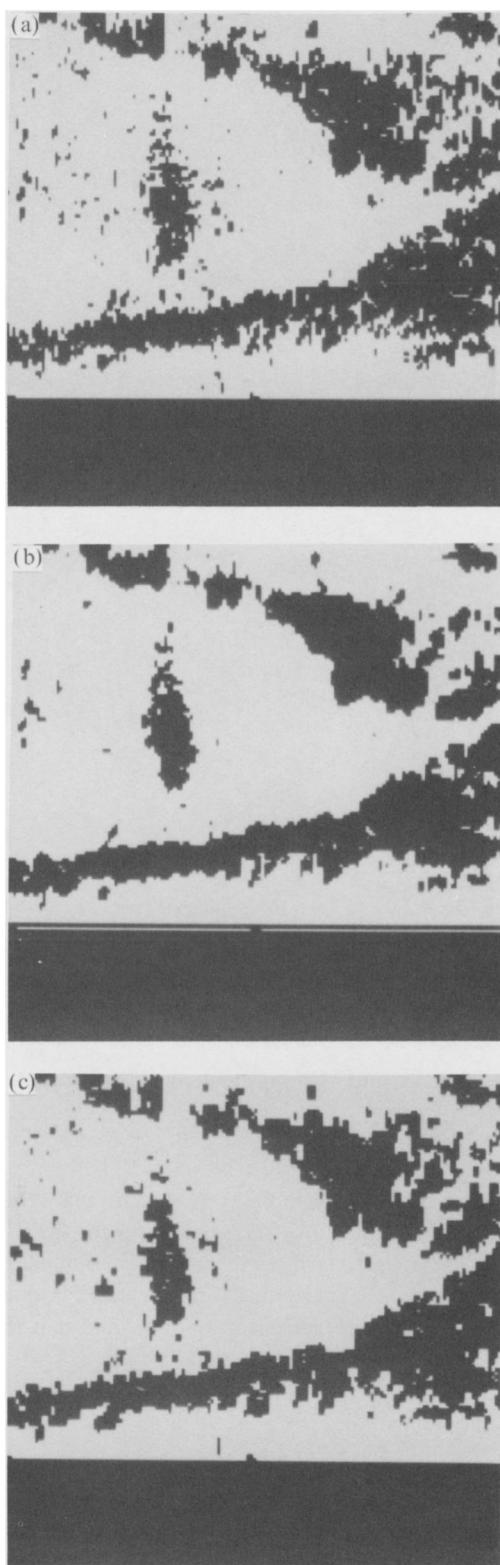


Fig. 4. Same image as in Fig. 3, (a) after applying the Besag noise reduction method; (b) after applying the relaxation noise reduction method.

where $H^l(c, V_{ijk})$ is the number of pixels in the three-dimensional neighbourhood whose class labels agree with c at iteration l . The initiation was done as for the relaxation procedure, and the value $\beta = 0.5$ was found to be satisfactory. Iterated conditional modes were found with ten iterations. The resulting noise reduced image for one index in time is displayed in Fig. 4. It can be seen that some of the noise has been smoothed and that the edges are reasonably well reproduced.

The binarisation was performed by a locally adaptive thresholding algorithm. The results from applying this to the original and the noise reduced images in Fig. 4 are shown in Fig. 5. Taxt *et al.* (1990) judge the relaxation approach to be superior with respect to quality in the binary image. Based on these binary images additional cleaning by removing binary objects with small temporal extent is performed. The boundary can be determined from this cleaned image, see Storvik (1992).



This multi-stage one task at a time approach to problem solving seems to be representative for many current applications in applied image analysis. It is not necessarily the best approach, however. An alternative approach to the problem would be to address the volume estimation directly from the initial images. Robust estimators may be able to reduce the influence of the noise. In a stepwise procedure the danger of filtering out important information concerning the main objective of the study is severe, and could lead to bias and a loss in efficiency. The "directly to the heart" approach looks attractive in that respect. The fact that it is the dynamic characteristics of the heart which are of interest calls for models for which smoothing along the time dimension is not too severe. A separate treatment of the time dimension seems necessary. See Storvik (1992).

4.3. Simulation of fault zones (Omre *et al.* 1992a, b)

Petroleum reservoirs in the North Sea are characterised by numerous fractures and faults. This is expected to have considerable impact on the production potential. Not merely the presence of faults but also the characteristics of the fault zone will have influence. The fault could slip along one continuous plane, or it may consist of a swarm of minor fractures. The exact flow mechanism across faults is not well understood at present. In this application a stochastic model for the break pattern is established. One fault zone will be considered and several realisations are generated. Each realisation is later taken as input to a fluid flow simulator in order to evaluate the impact of the break pattern on production.

The large faults, i.e. with offset above 20 m, can be observed on the seismic maps. They are characterised by centre line, vertical offset and lateral extension, see Fig. 6. The smaller fractures constituting the break pattern in the fault zone are below seismic resolution, however. It is these smaller fractures which will be modelled in this study. The stochastic model is based on general structural geological knowledge. The realisations generated from the model have to be constrained by the centre line, vertical offset and lateral extensions actually observed.

The stochastic model is based on the theory of marked point processes as outlined in sections 3.3 and 3.4.8. Each fracture is defined as a marked point $m = (x, \phi, \omega, \rho, \theta)$, see Fig. 7, in which x is the two-dimensional geographical reference point, a stochastic variable defined in the fault zone; ϕ is the orientation of fracture, constrained to be parallel to the

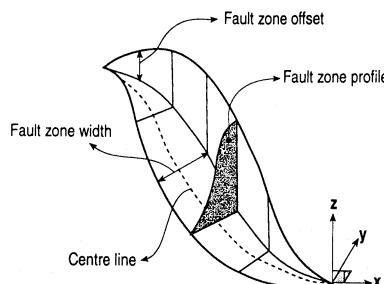


Fig. 6. Parameterisation of a fault zone.

Fig. 5. Same image as in Fig. 3, (a) after binarisation of original image; (b) after binarisation of relaxation method noise reduced image (b); (c) after binarisation of Besag method noise reduced image 4(a).

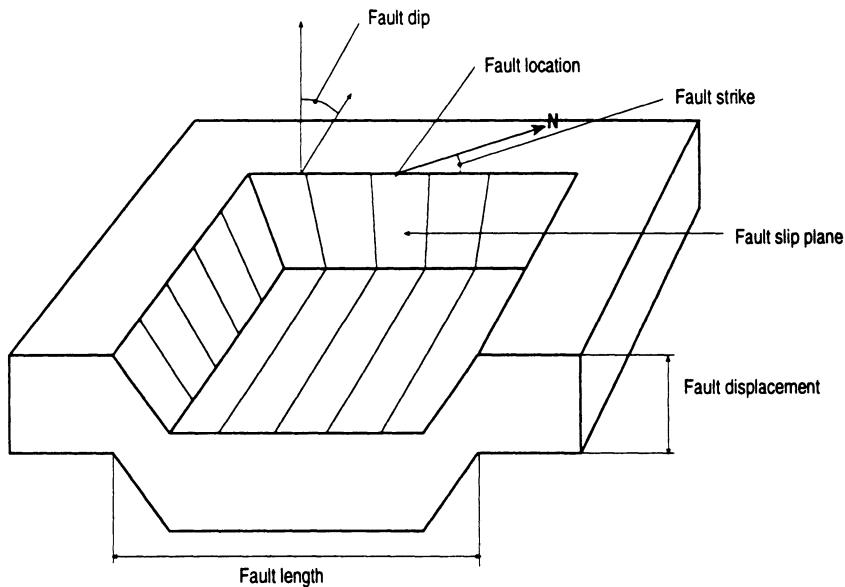


Fig. 7. Parameterisation of an individual fault.

centre line of the fault zone; ω is the stochastic width of fracture; ρ is the stochastic offset of fracture; and θ is the constant dip angle, as specified by user. The models applied in the present study specify that the joint probability for n fractures m_1, \dots, m_n is of the general form

$$f_n(m_1, \dots, m_n) = \text{const} \exp \left\{ - \sum_{i=1}^n b(m_i) - \sum_{i < j} c(m_i, m_j) \right\} \exp \{ -\sigma \Delta(m_1, \dots, m_n) \}. \quad (4.6)$$

Here $b(m)$ takes care of the dependence structure for attributes in each marked point, $c(m_i, m_j)$ models the pairwise interaction between marked points, and $\Delta(m_1, \dots, m_n)$ is finally a measure of deviation between the realisation (m_1, \dots, m_n) and some desired property, with accompanying strength parameter σ . Note the similarity to a technique mentioned in section 3.2.4. If exact or approximate constraints of the simulated marked points are needed then such are built into the deviance measure.

The model actually employed in our study used

$$b(m_i) = b_1(x_i) + b_2(\omega_i | x_i) + b_3(\rho_i | \omega_i), \quad (4.7)$$

in which $b_1(x)$ is a function defining the fracture frequency in the fault zone dependent upon the distance from x to the centre line of the fault; $b_2(\omega | x)$ is another function defined in the fault zone where the width can be a function of the distance from x to the centre line of the fault; and $b_3(\rho | \omega)$ is a function relating the width of the fracture to the offset, taken as $(\rho - \mu\omega^2)^2 / (2(\omega^2\eta)^2)$ for certain parameters μ and η . Note that steeper slopes in some areas of the fault zone can be realised by either higher frequencies of fractures or larger expected width of each fracture which in turn is correlated with offset. The trade-off between these two effects is governed by a parameter α . Furthermore $c(x_i, x_j)$ was taken as $c_0(x_i - x_j)$, where

$$c_0(\Delta x) = \begin{cases} k(\phi)/|\Delta x|^2 & \text{for } 0 \leq |\Delta x| \leq x_0 k(\phi)^{1/2}, \\ 1/x_0^2 & \text{for } |\Delta x| \geq x_0 k(\phi)^{1/2}, \end{cases}$$

with $(|\Delta x|, k(\phi))$ being the polar coordinates for Δx , and $k(\phi)$ representing an elliptical anisotropy factor. Finally the $\Delta(m_1, \dots, m_n)$ measure of deviance used, to be scaled with σ afterwards, is

$$\Delta(m_1, \dots, m_n) = \frac{1}{|L|} \int_L |(\text{total offset at } u) - (\text{realised offset at } u \text{ by } \{m_1, \dots, m_n\})| du,$$

with L being the centre line of fault and $|L|$ its length.

In the study the actual model was explored by simulation in order to see if it was suited for generating realistic realisations of break patterns in faults, using the Ripley–Kelly algorithm, see section 3.3.2. In this summary the sensitivity to $b_1(x)$ and $b_2(\omega | x)$ will be reported on. In Figs 8 and 9 the interaction function $c_0(\cdot)$ is used, the strength parameter σ is assigned a relatively large value, and the effect of having steeper slope in the middle of the fault area is imposed. In the figure the upper display is a bird's eye view and the locations of the fractures are exposed. Note their "repulsive" appearance. The middle display corresponds to facing the fault and the respective offsets can be observed. Note that the total offset in the fault is reproduced since global constraints are imposed. The lower display shows the average slope profile for the fault in solid line and three profiles at arbitrary

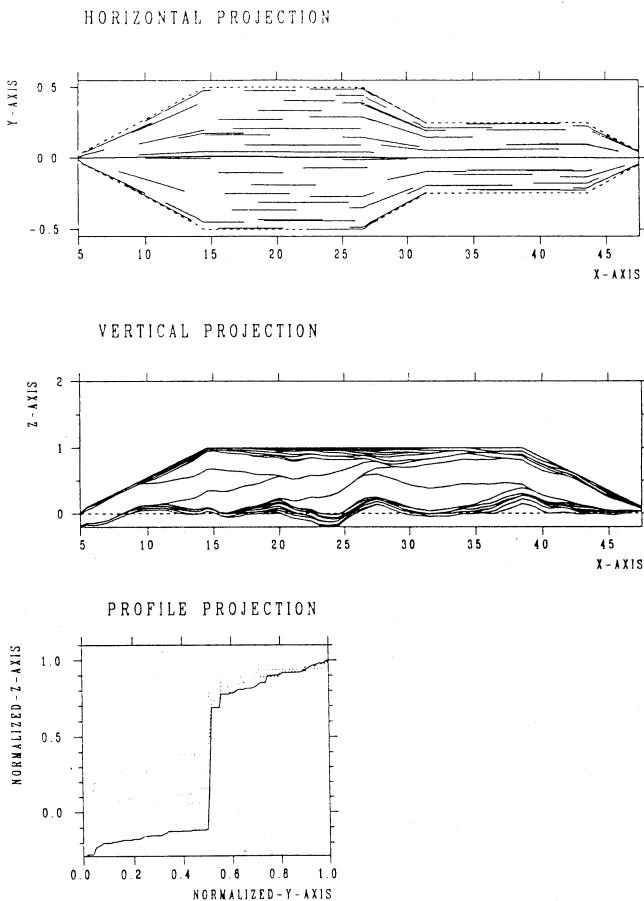


Fig. 8. Realisation of a heterogeneous fault zone. The fault zone offset is reproduced; required steepness close to centre line is realised by larger displacement in individual faults.

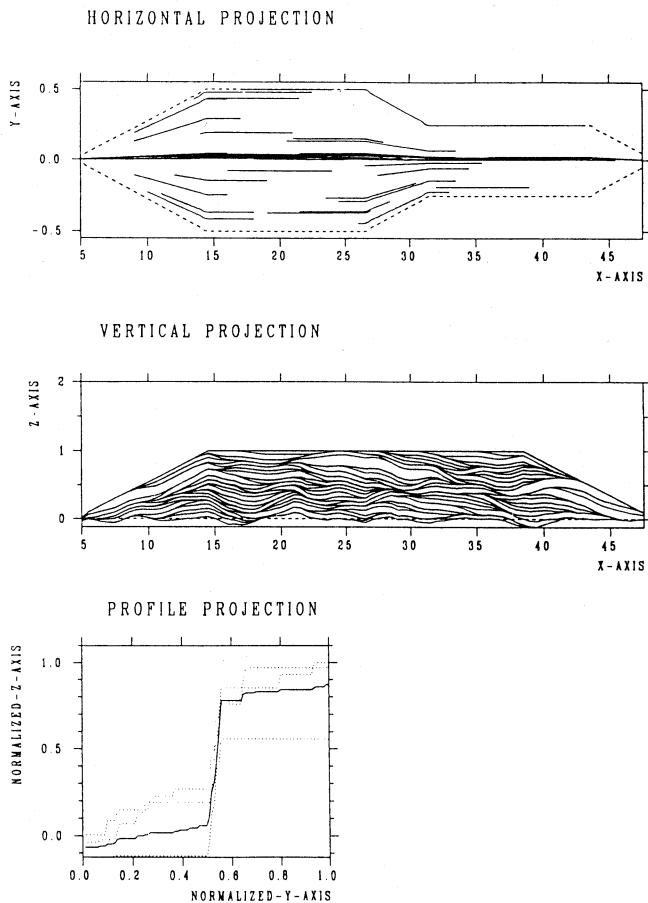


Fig. 9. Realisation of a heterogeneous fault zone. The fault zone offset is reproduced; required steepness close to centre line is realised by higher intensity of individual faults.

locations in dotted lines. In Fig. 8 the steepness is realised by having larger width and offset in the centre, while in Fig. 9 it is realised by having higher frequency of fractures in the centre. It is also possible to use a trade-off between the two.

The geologists have approved the results, and the study has since proceeded towards evaluation of fluid flow across the fault zone. Further methodological work will aim at understanding the interaction between the parameters when enforcing global constraints. Parameter estimation from observations of fractures in wells and from comparable outcrop data will also be studied.

4.4. Spatial prediction of air pollution from space-time observations (Høst *et al.*, 1994)

The awareness of the possible consequences of changes in the environment has contributed to increasing interest in pollution monitoring. Control of international agreements on reduction of emissions to the atmosphere will require careful air pollution monitoring, supplemented with a thorough statistical evaluation. Air pollution will normally be modelled as a spatial-temporal phenomenon, and the available data take the form of time series in a

number of fixed locations. Challenging problems like spatial interpolation and evaluation of temporal trends in arbitrary locations remain mostly unstudied.

In the study reported on here the problem of spatial interpolation at a given time point is addressed. The air pollution data used are collected by the European Monitoring Evaluation Programme (EMEP). The variable considered is sulphur concentration in units of micrograms per cubic meter, $\mu\text{g}/\text{m}^3$. The data are presently collected in more than one hundred locations all over Europe, but in this study data from the six years 1980–1985 are used, in the form of monthly averages of sulphur concentrations in 42 fixed locations. The locations provide a good areal coverage over Northern and Central Europe. The time series are log-transformed and are denoted $\{y(x_i, t_j) : i \leq 42, j \leq 72\}$, with x and t denoting geographical position and time respectively. The time series for the monitoring locations expose sizeable spatial differences as well as seasonal variations.

The statistical objective we focus on is to estimate the surface $y(x, t)$, for x in the domain \mathcal{D} , for a given t among \mathcal{T} , the 72 time points used. A natural model for this purpose is

$$Y(x, t) = \eta(t) + M(x) + V(x, t), \quad x \in \mathcal{D}; t \in \mathcal{T}, \quad (4.8)$$

where $\eta(t)$ is centred temporal drift, containing seasonal variations but independent of location, $M(x)$ is the spatial drift independent of time, and $V(x, t)$ is the centred space–time residual. We choose to represent the latter as $V(x, t) = S(x, t)U(x, t)$, where $S(x, t)$ is residual standard deviation and $U(x, t)$ is the normalised residual. Moreover, the residual standard deviation is factorised into a temporal and a spatial term, $S(x, t) = \kappa(t)H(x)$. The temporal drift is represented deterministically since its variation is considered to contribute only marginally to the uncertainty of the final interpolator. If the aim of the study had been temporal forecasting it would have been necessary to model the $\eta(t)$ function stochastically as well.

The model is defined up to second order and employs the following parameters: spatial drift is based on an ordinary Kriging model, with

$$EM(x) = \mu_M, \quad \text{var}\{M(x) - M(x')\} = 2\gamma_M(x - x').$$

The residual standard deviation is similarly based on an ordinary Kriging model, with

$$EH(x) = \mu_H, \quad \text{var}\{H(x) - H(x')\} = 2\sigma_H^2(1 - \rho_H(x - x')).$$

Finally the normalised residual is modelled with a location dependent correlation function, discussed in Switzer (1989) and Loader & Switzer (1992), having

$$EU(x, t) = 0, \quad \text{var}\{U(x, t) - U(x', t)\} = 2\{1 - \rho_U(x, x')\}.$$

The model differs from what is traditionally being used in that the standard deviation is spatially varying and that the spatial correlation function in the residuals is location dependent. The correlation functions and parameters of the model can be estimated from the $y(x_i, t_j)$ -data. See Høst *et al.* (1994) for development of some reasonable but not necessarily optimal estimators. It is difficult to establish exact properties of these, but they are based on a fair amount of data and should be sufficiently reliable.

The spatial interpolator for an arbitrary location x_0 , at the time point t under consideration, is

$$Y^*(x_0, t) = \hat{\eta}(t) + M^*(x_0) + \hat{\kappa}(t)H^*(x_0)U^*(x_0, t). \quad (4.9)$$

Here $\hat{\eta}(\cdot)$ and $\hat{\kappa}(\cdot)$ are estimates of temporal drift based on a simple smoother on the available data, $M^*(\cdot)$ and $H^*(\cdot)$ are ordinary Kriging predictors, and $U^*(\cdot, \cdot)$ is determined by the procedure for spatial interpolation discussed in Switzer (1989) and Loader & Switzer (1992). In the end the predicted sulphur concentration is obtained by the inverse log-transform.

The interpolator Y^* is not optimal in the mean square error sense since the three components M, H, U are predicted independently. Scatter plots have indeed indicated near independence for the situation at hand. It is worth noting that the prediction variance of (4.9) can be calculated explicitly for the case of independent M, H, U , and can be expressed as

$$\begin{aligned}\text{var} \{ Y(x_0, t) - Y^*(x_0, t) \} &= \text{var} \{ M(x_0) - M^*(x_0) \} + \kappa(t)^2 \mu_H \text{var} \{ U(x_0, t) - U^*(x_0, t) \} \\ &\quad + \kappa(t)^2 \sigma_H^2 [1 - \text{corr} \{ H(x_0), H^*(x_0) \} \text{corr} \{ U(x_0, t), U^*(x_0, t) \}].\end{aligned}$$

The interpolator (4.9) and its prediction variance have the exactness property, in the sense that observations are correctly predicted with prediction variance zero at the data locations. In general, one will obtain weights associated with the observations which are location specific and not only dependent on the configuration of data locations. This makes the interpolator different from the simple and ordinary Kriging interpolator methods. Furthermore, the model provides more detailed and possibly more accurate representation of interpolation errors than traditional Kriging.

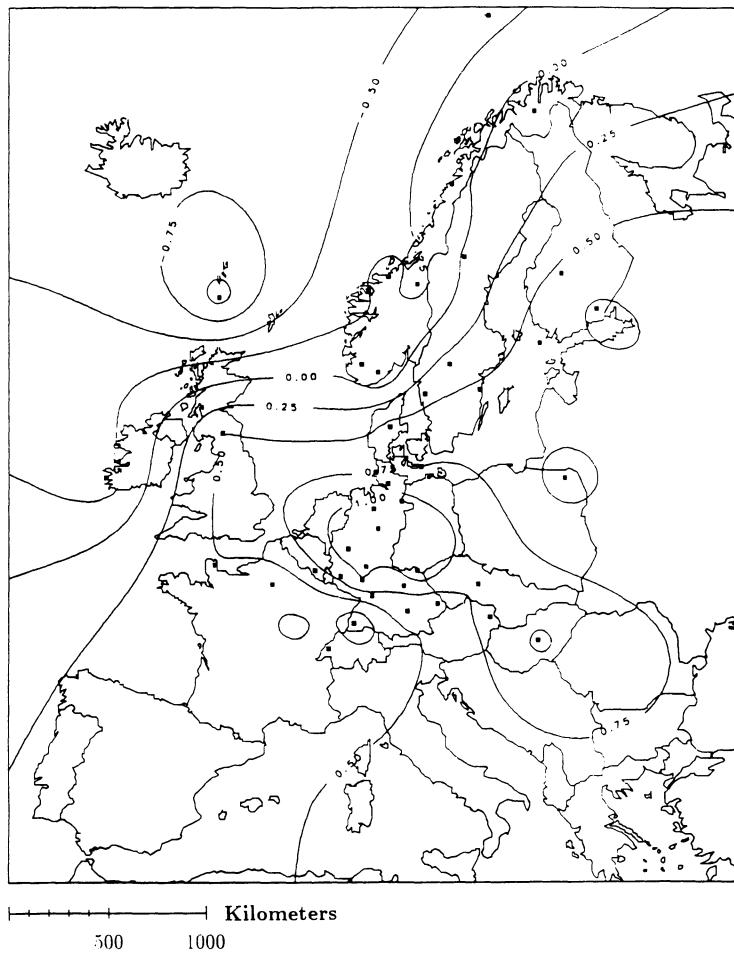


Fig. 10. (a)

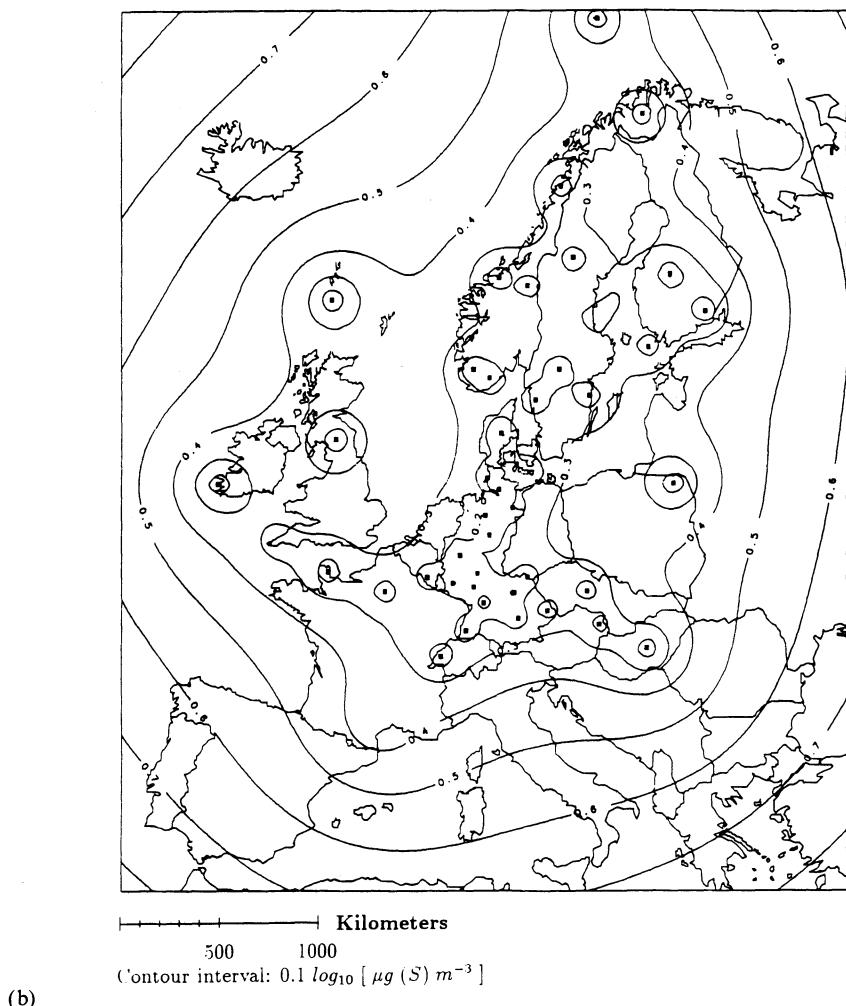


Fig. 10. (a) Interpolated log sulphur concentration January 1984, along with (b) its estimated standard deviation surface.

The illustration shows the interpolated log sulphur concentration with error bounds over Europe for January 1984. The prediction is presented on a grid of size 375×450 over Europe. Figure 10 presents contour maps of the interpolated log sulphur concentration and the associated interpolation error map. The monitoring locations are identified by dots.

A comparison with traditional Kriging techniques, discussed in section 3.1.2, has been carried out via cross-validation methods, and has shown that both bias and variance were reduced when applying the model and methods described here. Further work on evaluation of temporal trends in the average of sulphur concentration and other air quality variables over given regions is being pursued.

5. Closing remarks

There is an increasing interest in spatial and spatial-temporal statistical methodology. Vast amounts of automatically collected data have made new applications accessible for spatial

evaluation. Inexpensive, fast-processing computers with graphical capabilities are also becoming more easily available, and provide the technological tools for most spatial statistical analysis. There is also a trend towards larger emphasis on decision-making under uncertainty, and one might predict a shift in emphasis from computer science to modelling. Our experience with and expectations for spatial and spatial-temporal statistics can tentatively be summarised as follows.

There is a wide variety of challenging applications. User groups are normally positive since they are often poorly trained in handling spatial data and realise that their evaluation is insufficient. The possibility of including user experience through "prior guesses" is highly appreciated. Numerical results can often be supplemented with graphical displays, and this simplifies verification and interaction with the users. The main experience is, however, that *every problem is unique*. This is true even more so than in traditional statistics because of the modelling possibilities and the large variety of sampling designs.

One of the things to note on the methodological side is that many natural models do not admit explicit closed form expression solutions. The mathematical complexity increases considerably for higher dimensions in the reference space. The intractability of Markov properties of Gaussian random functions is but one example. Efforts should be made to obtain further theoretical and analytical results. To compensate for the lack of explicit results stochastic simulation is being used extensively. It has in many cases proved successful, and valuable insights into the models have been reached. For surprisingly many models, however, simulation techniques have proved to be unreliable or too time-consuming. The disappointingly slow convergence of the Metropolis algorithm and Gibbs sampler for simulation of Markov random fields are examples of this. Further research on new model formulations with associated reliable and efficient simulation algorithms are certainly needed. The availability of vector and parallel processing computers should be taken to advantage.

For several of the most often used models there is a lack of reliable estimators for the model parameters. Constructing such estimators is a difficult task, since the parameters are often inter-related and the sampling designs vary considerably. Criteria for selection between different models are rather undeveloped and research along such lines should be initiated. Model adaptation procedures should be cross-validated and resampling methods should be further developed for spatial models.

As can be seen, despite recent and healthy progress in the field, there seem to be more questions than answers concerning spatial and spatial-temporal statistics so far. These fields will provide challenges for both theoretical and applied statisticians for the years to come.

Acknowledgements

We have been privileged to work in a stimulating environment at the Norwegian Computing Centre, with fellow statisticians of high standard as well as with demanding clients from several sectors of industry and research. Our work over the years has partly been supported by the Royal Norwegian Council for Scientific and Technological Research. Finally we are grateful to the programme committee of the Nordic Conference in Odense 1990 for inviting us to give the forum lectures which formed the core of the present article, and to Editor Elja Arjas for encouragement and patience.

References

- Abrahamsen, P. (1993). Bayesian Kriging for seismic depth conversion of a multi-layer reservoir. In *Geostatistics Troia '92*, Vol. I (ed. A. Soares), 385–398. Kluwer Academic, Amsterdam.

- Abrahamsen, P., Omre, H. & Lia, O. (1991). Stochastic models for seismic depth conversion of geological horizons. *Soc. Petroleum Engrs* **23**, 329–341.
- Adler, R. J. (1981). *The geometry of random fields*. Wiley, New York.
- Arak, T. & Surgailis, D. (1989). Markov fields with polygonal realisations. *Probab. Theory Rel. Fields* **80**, 543–580.
- Arak, T., Clifford, P. & Surgailis, D. (1994). Markov random graphs. *Adv. Applied Probab.* in press.
- Baddeley, A. & Møller, J. (1989). Nearest-neighbour Markov point processes and random sets. *Int. Statist. Rev.* **57**, 89–121.
- Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems [with discussion]. *J. Roy. Statist. Soc. B* **36**, 192–236.
- Besag, J. (1986). On the statistical analysis of dirty pictures [with discussion]. *J. Roy. Statist. Soc. B* **48**, 259–302.
- Box, G. E. F. & Jenkins, G. (1976). *Time series analysis: forecasting and control*. Holden-Day, Johannesburg.
- Bølviken, E. (1993). Statistical methods using regime processes and other non-linear state space methods. Statistical research report, University of Oslo.
- Bølviken, E. & Helgeland, J. (1989). Some models and algorithms in segmented time series analysis. Technical report SAND/89, Norwegian Computing Centre, Oslo.
- Bølviken, E., Helgeland, J. & Storvik, G. (1991). Stochastic regime models and wireline log data. *Bull. Int. Statist. Inst.*, 48th Biennial Session, Cairo 1991.
- Christensen, R. (1990). The equivalence of predictions from universal Kriging and intrinsic random-function Kriging. *Math. Geol.* **22**, 655–664.
- Clemetsen, R., Hurst, A., Omre, H. & Knarud, H. (1989). A computer program for evaluation of fluvial reservoirs. *Proceedings 2nd international conference on North Sea oil and gas reservoirs* (ed. J. Kleppe), 373–385. Graham & Trotham, Norwegian Institute of Technology, Trondheim.
- Clifford, P. & Middleton, R. D. (1989). Reconstruction of polygonal images. *J. Appl. Statist.* **16**, 409–422.
- Comets, F. & Gidas, B. (1992). Parameter estimation for Gibbs distributions from partially observed data. *Ann. Appl. Probab.* **2**, 142–170.
- Conradsen, K. & Nielsen, B. K. (1987). Classification of digital images. *Proceedings 2nd int. Tampere Conf. in Statistics*.
- Cressie, N. (1991). *Statistics for spatial data*. Wiley, New York.
- Davis, J. C. (1973). *Statistics and data analysis in geology*. Wiley, New York.
- Diggle, P. J. (1979). On parameter estimation and goodness-of-fit testing for spatial point patterns. *Biometrics* **35**, 87–101.
- Diggle, P. J. (1983). *Statistical analysis of spatial point patterns*. Academic Press, London.
- Diggle, P. J., Gates, D. J. & Stallard, A. (1987). A nonparametric estimator for pairwise interaction processes. *Biometrika* **74**, 763–770.
- Esbensen, K. & Geladi, P. (1989). Strategy of multivariate image analysis (MIA). *Chem. Intell. Lab. Syst.* **7**, 67–86.
- Feder, J. (1988). *Fractals*. Plenum Press, New York.
- Fiksel, T. (1988). Estimation of interaction potentials of Gibbsian point processes. *Statistics* **19**, 77–80.
- Fiskum, S. (1986). Relaxation methods used in supervised classification of multispectral data [in Norwegian]. Report No. 788. Norwegian Computing Centre, Oslo.
- Geman, S. & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Machine Intell.* **6**, 721–741.
- Georgsen, F. & Omre, H. (1993). Combining fibre processes and Gaussian random functions for modelling fluvial reservoirs. In *Geostatistics Troia '92*, Vol. I (ed. A. Soares), 425–440. Kluwer Academic, Amsterdam.
- Gidas, B. (1985). Nonstationary Markov chains and convergence of the annealing algorithm. *J. Statist. Phys.* **39**, 73–131.
- Gidas, B. (1991). Parameter estimation for Gibbs distributions. I: fully observed data. In *Markov random fields: theory and applications* (eds. R. Chellappa & R. Jain). Academic Press, New York.
- Godtliebsen, F. (1989). A study of image improvement techniques applied to NMR images. Dr. Ing. thesis, Norwegian Institute of Technology, Trondheim.
- Grabarnik, P. & Särkkä, A. (1992). On parameter estimation of marked Gibbs point process. Technical report, Department of Statistics, University of Jyväskylä.
- Green, P. J. (1986). Discussion contribution to Besag's 'Statistical analysis of dirty pictures', *J. Roy. Statist. Soc. B* **48**, 284–285.

- Green, P. J. & Titterington, D. M. (1987). Recursive methods in image processing. *Bull. Int. Statist. Inst.*, invited paper in the session on spatial image analysis, Tokyo, 51–67.
- Greig, D. M., Porteous, B. T. & Seheult, A. H. (1989). Exact MAP estimation for binary images. *J. Roy. Statist. Soc. 51*, 271–279.
- Hall, P. G. (1988). *Introduction to the theory of coverage processes*. Wiley, New York.
- Halvorsen, K. & Strand, G.-H. (1987). Geostatistical analysis of pollution data: a study of methods. Technical report STAT/87, Norwegian Computing Centre, Oslo.
- Haslett, J. (1985). Maximum likelihood discriminant analysis on the plane using a Markovian model of spatial context. *Pattern Recog. 18*, 287–296.
- Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika 57*, 97–109.
- Helgeland, J., Hjort, N. L. & Sæbø, H. V. (1984). Spatial sampling strategies: a review and some possible applications. Technical report, Norwegian Computing Centre, Oslo.
- Hjort, J. (1914). Fluctuations in the great fisheries of Northern Europe viewed in the light of biological research. *Conseil permanent international pour l'exploration de la mer: Rap. Proc. Verbaux 20*.
- Hjort, J. & Murray, J. (1912). *The depths of the oceans*. Cambridge University Press, London.
- Hjort, N. L. (1985a). Neighbourhood based classification of remotely sensed data based on geometric probability models. Technical report 10/NSF, Department of Statistics, Stanford University.
- Hjort, N. L. (1985b). Estimating parameters in neighbourhood based classifiers for remotely sensed data, using unclassified vectors. Technical report 12/NSF, Department of Statistics, Stanford University.
- Hjort, N. L. (1986). *Statistical symbol recognition*. Research monograph, Norwegian Computing Centre, Oslo.
- Hjort, N. L. (1987). Classification of 2×2 blocks of pixels in image analysis. Technical report BILD/87, Norwegian Computing Centre, Oslo.
- Hjort, N. L. (1993). A quasi-likelihood method for estimating parameters in spatial covariance functions. Technical report SAND/93, Norwegian Computing Centre, Oslo.
- Hjort, N. L. & Mohn, E. (1984). A comparison of some contextual methods in remote sensing. *Proceedings 18th international symposium on remote sensing of the environment*, CNES, Paris, 1693–1702.
- Hjort, N. L. & Mohn, E. (1985). On the contextual classification of data from high resolution satellites. *Proceedings 4th Scandinavian conference on image analysis*, 391–399. Tapir, Trondheim.
- Hjort, N. L. & Mohn, E. (1987). Topics in the statistical analysis of remotely sensed data. *Bull. Int. Statist. Inst.*, invited paper in the session on spatial image analysis, Tokyo, 23–47.
- Hjort, N. L. & Taxt, T. (1988). Automatic training in statistical pattern recognition. *Proceedings int. conference on pattern recognition*, Sicily 1987.
- Hjort, N. L., Mohn, E. & Storvik, G. (1987). A simulation study of some contextual classification methods for remotely sensed data. *IEEE Trans. Remote Sensing Geosci. 15*, 796–804.
- Hjort, N. L., Holden, L. & Omre, H. (1989). Modelling sedimentary facies by Markov random fields. Technical report SAND/89, Norwegian Computing Centre, Oslo.
- Holden, L. & Tjelmeland, H. (1990). A program system for simulation of reservoir architecture and reservoir properties. Technical report SAND/90, Norwegian Computing Centre, Oslo.
- Homleid, M. (1992). TOVS data processing and the potential use of the data in numerical weather prediction models at the Norwegian Meteorological Institute. Technical report No. 83, Norwegian Meteorological Institute, Oslo.
- Høiberg, J., Omre, H. & Tjelmeland, H. (1990). Large scale barriers in extensively drilled reservoirs. *Proceedings 2nd European Conference on the Mathematics of Oil Recovery*, Arles, France.
- Høiberg, J., Omre, H. & Tjelmeland, H. (1992). A stochastic model for shale distribution in petroleum reservoirs. *Proceedings 2nd Codata conference on geomathematics and geostatistics, Sciences de la Terre. Sér. Inf. Nancy 31*, 135–153.
- Høst, G., Omre, H. & Sæbø, H. V. (1989). Combining field observations and remotely sensed data in pollution monitoring: a spatial statistical approach. Technical report STAT/89, Norwegian Computing Centre, Oslo.
- Høst, G., Omre, H. & Switzer, P. (1994). Spatial interpolation errors for monitoring data. *J. Amer. Statist. Assoc. to appear*.
- Isaaks, E. H. and Srivastava, R. M. (1989). *An introduction to applied geostatistics*. Oxford University Press, New York.
- Jensen, J. L. & Møller, J. (1991). Pseudolikelihood for exponential family models of spatial point processes. *Ann. Appl. Probab. 1*, 445–461.
- Journel, A. G. (1983). Nonparametric estimation of spatial distributions. *Math. Geology 15*, 445–468.

- Journel, A. G. (1986). *Fundamentals of geostatistics in five lessons*. American Geophysical Union, Washington.
- Journel, A. G. & Huijbregts, C. J. (1978). *Mining geostatistics*. Academic Press, New York.
- Kay, J. W. & Titterington, D. M. (1986). Image labelling and the statistical analysis of incomplete data. *Proceedings 2nd international conference on image processing and its applications*, IEEE, Imperial College of Science and Technology, London.
- Künsch, H. R. (1986). Discussion contribution to Besag's 'Statistical analysis of dirty pictures', *J. Roy. Statist. Soc. B* **48**, 294.
- Lakshmanan, S. & Derin, H. (1989). Simultaneous parameter estimation and segmentation of Gibbs random fields using simulated annealing. *IEEE Trans. Pattern Anal. Machine Intell.* **11**, 799–813.
- Le, N. D. & Zidek, J. V. (1992). Interpolation with uncertain spatial covariances: a Bayesian alternative to Kriging. *J. Multivar. Analysis* **43**, 351–374.
- Loader, C. & Switzer, P. (1992). Spatial covariance estimation for monitoring data. In *Statistics in the environmental & earth sciences* (eds A. T. Walden & P. Guttorp). Edward Arnold, London.
- Lotwick, H. W. & Silverman, B. W. (1981). Convergence of spatial birth-and-death processes. *Math. Proc. Camb. Phil. Soc.* **90**, 155–165.
- Lundervold, A., Moen, K. & Taxt, T. (1988). Automatic recognition of normal and pathological tissue types in NMR-images: a feasibility study using contextual classification methods. *Proceedings NOBIM conference*, Oslo. Report No. 818, Norwegian Computing Centre, Oslo, 189–192.
- Mardia, K. V. & Marshall, R. J. (1984). Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika* **71**, 135–146.
- Mardia, K. V. and Watkins, A. J. (1989). On multimodality of the likelihood in the spatial linear model. *Biometrika* **76**, 289–295.
- Marroquin, J., Mitter, S. & Poggio, T. (1987). Probabilistic solution of ill-posed problems in computational vision. *J. Amer. Statist. Assoc.* **82**, 76–89.
- Matérn, B. (1960). Spatial variation. *Meddelanden från Statens Skogforskningsinstitut* **49**, 1–144. Reprinted 1986 in *Lect. Notes Statist.* **36**, Springer-Verlag, Berlin.
- Matheron, G. (1965). *Les variables regionalisees et leur estimation*. Masson & Cie, Paris.
- Matheron, G. (1973). The intrinsic random functions and their applications. *Adv. Appl. Probab.* **5**, 439–468.
- Matheron, G. (1976). A simple substitute for conditional expectation: disjunctive Kriging. *Proceedings 1st NATO geostatistical conference* (eds. Fabbri *et al.*), 221–236. Reidel, Dordrecht.
- Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. (1953). Equations of state calculations by fast computing machines. *J. Chem. Phys.* **21**, 1087–1092.
- Moyeed, R. A. & Baddeley, A. J. (1991). Stochastic approximation of the MLE for a spatial point pattern. *Scand. J. Statist.* **18**, 39–50.
- Nicholls, G. K. & Petrou, M. (1993). A generalisation of renormalisation group methods for multi resolution image analysis. Technical report, Department of Electronic and Electrical Engineering, University of Surrey.
- Omre, H. (1984). Alternative variogram estimators in geostatistics. Ph.D. thesis, Earth Sciences Department, Stanford University.
- Omre, H. (1987). Bayesian Kriging—merging observations and qualified guesses in Kriging. *Math. Geol.* **19**, 25–39.
- Omre, H. (1992). Stochastic models for reservoir characterisation. In *Recent advances in improved oil recovery methods for North Sea sandstone reservoirs* (eds. S. M. Skjæveland & J. Kleppe), 141–149. Norwegian Petroleum Directorate, Stavanger.
- Omre, H. & Halvorsen, K. (1989). The Bayesian bridge between simple and universal Kriging. *Math. Geol.* **21**, 767–786.
- Omre, H. & Solna, K. (1991a). Uncertainty assessment of fish biomass prediction. Technical report SAND/91, Norwegian Computing Centre, Oslo.
- Omre, H. & Solna, K. (1991b). Relationship between spawning potential and length for Barents Sea capelin. Technical report SAND/91. Norwegian Computer Centre, Oslo.
- Omre, H., Halvorsen, K. & Berteig, V. (1989). A Bayesian approach to Kriging. *Geostatistics I* (ed. M. Armstrong), 109–126. Kluwer Academic, Amsterdam.
- Omre, H., Sølna, K. & Tørudbakken, B. (1992a). Stochastic modelling and simulation of fault zones. *Proceedings 2nd Codata conference on geomathematics and geostatistics, sciences de la terre, Sér. Inf. Nancy* **31**, 177–194.
- Omre, H., Sølna, K., Dahl, N. & Tørudbakken, B. (1992b). Impact of fault heterogeneity in fault zones on fluid flow. *Proceedings 3rd International Conference on North Sea Oil and Gas Reservoirs*.

- Omre, H., Sølna, K. & Tjelmeland, H. (1993). Simulation of random functions on large lattices. In *Geostatistics Troia '92*, Vol. I (ed. A. Soares), 179–200. Kluwer Academic, Amsterdam.
- Owen, A. (1984). A neighbourhood-based classifier for LANDSAT data. *Canad. J. Statist.* **12**, 191–200.
- Owen, A. (1989). Image segmentation via iterated conditional expectations. Technical report No. 254, Department of Statistics, University of Chicago.
- Pickard, D. K. (1977). A curious binary lattice process. *J. Appl. Probab.* **14**, 717–731.
- Pickard, D. K. (1987). Inference for discrete Markov random fields: the simplest nontrivial case. *J. Amer. Statist. Assoc.* **82**, 90–96.
- Pilz, J. (1991). *Bayesian estimation and experimental design in linear regression models*. Wiley, New York.
- Pripp, P. O. (1990). Binary Markov random fields applied to automatic recognition of hand-written symbols [in Norwegian]. Graduate thesis, Department of Mathematics, University of Oslo.
- Ripley, B. D. (1977). Modelling spatial patterns [with discussion]. *J. Roy. Statist. Soc. B* **39**, 172–212.
- Ripley, B. D. (1981). *Spatial statistics*. Wiley, New York.
- Ripley, B. D. (1986). Statistics, images and pattern recognition [with discussion]. *Canad. J. Statist.* **14**, 83–11.
- Ripley, B. D. (1987). *Stochastic simulation*. Wiley, New York.
- Ripley, B. D. (1988). *Statistical inference for spatial processes*. Cambridge University Press, Cambridge.
- Ripley, B. D. (1989a). Gibbsian interaction models. In *Spatial statistics: past, present and future* (ed. D. A. Griffiths). Image, New York.
- Ripley, B. D. (1989b). The uses of spatial models as image priors. In *Spatial statistics & imaging* (ed. A. Possolo), *IMS Lecture Notes*, 29 pp.
- Ripley, B. D. (1992). Stochastic models for the distributions of rock types in petroleum reservoirs. In *Statistics in the environmental & earth sciences* (eds A. T. Walden & P. Guttorp). Edward Arnold, London.
- Ripley, B. D. & Kelly, F. P. (1977). Markov point processes. *J. Lond. Math. Soc.* **15**, 188–192.
- Sampson, P. D. & Guttorp, P. (1992). Nonparametric estimation of nonstationary spatial covariance structure. *J. Amer. Statist. Assoc.* **87**, 108–119.
- Sandjiviy, L. (1984). The factorial Kriging analysis of regionalised data. Its application to geochemical prospecting. *Geostatistics for natural resources characterisation I* (eds G. Verly, M. David, A. G. Journel & A. Marechal), 559–571. Reidel, Dordrecht.
- Särkkä, A. (1993). The pseudo-likelihood approach for Gibbs point processes in connection with field observations. Technical report, University of Jyväskylä.
- Schweder, T. & Høst, G. (1991). Integrating experimental data and survey data through a simulation model to obtain estimates of $g(0)$. *Proceedings International Whaling Committee*, Cambridge.
- Schweder, T., Øien, N. & Høst, G. (1990). Estimates of the detection probability for shipboard surveys of Northeastern Atlantic minke whales, based on a parallel ship experiment. Technical report STAT/90, Norwegian Computing Centre, Oslo.
- Solow, A. R. (1990). Geostatistical cross-validation: a cautionary note. *Math. Geology* **22**, 637–639.
- Stein, M. (1987). Minimum norm quadratic estimation of spatial variograms. *J. Amer. Statist. Assoc.* **82**, 765–772.
- Stein, M. (1990a). Bounds on the efficiency of linear predictions using an incorrect covariance function. *Ann. Statist.* **18**, 1116–1138.
- Stein, M. (1990b). A comparison of generalised cross validation and modified maximum likelihood for estimating the parameters of a stochastic process. *Ann. Statist.* **18**, 1139–1157.
- Storvik, G. (1992). A Bayesian approach to dynamic contours. Doctoral thesis, Norwegian Computing Centre Report No. 860.
- Storvik, G. & Switzer, P. (1992). Space-time modelling of simply connected objects: an application to detection of left ventricular cardiac boundaries from ultrasound images. Technical report, Norwegian Computing Centre, Oslo.
- Stoyan, D., Kendall, W. S. & Mecke, J. (1987). *Stochastic geometry and its applications*. Akademie-Verlag, Berlin DDR.
- Strand, G.-H. (1989). Remote sensing of forest: an inductive approach. Report No. 824, Norwegian Computing Centre, Oslo.
- Switzer, P. (1965). A random set process in the plane with a Markov property. *Ann. Math. Statist.* **36**, 1859–1863.
- Switzer, P. (1984). Inference for spatial autocorrelation functions. *Geostatistics for natural resources characterization, I* (eds G. Verly, M. David, A. G. Journel & A. Marechal), 127–140. Reidel, Dordrecht.
- Switzer, P. (1989). Non-stationary spatial covariances estimated from monitoring data. In *Geostatistics I* (ed. M. Armstrong), 127–138. Kluwer Academic, Amsterdam.

- Switzer, P. & Green, A. A. (1984). Min/max autocorrelation factors for multivariate spatial imagery. Technical Report, Department of Statistics, Stanford University.
- Switzer, P. & Ingebritsen, S. E. (1986). Ordering of time-difference data from multispectral imagery. *Remote Sensing Environ.* **20**, 85–94.
- Sæbø, H. V., Bråten, K. H., Hjort, N. L., Llewellyn, B. & Mohn, E. (1985). Contextual classification of remotely sensed data: statistical methods and development of a system. Report No. 768, Norwegian Computing Centre, Oslo.
- Taxt, T. & Bølviken, E. (1991). Relaxation using models from quantum mechanics. *Pattern Recog.* **24**, 695–709.
- Taxt, T., Lundervold, A. & Angelsen, B. (1990). Noise reduction and segmentation in time-varying ultrasound images. *Proceedings 10th international conference on pattern recognition: systems and applications*, Atlantic City.
- Tjelmeland, H. & Holden, L. (1993). Semi-Markov random fields. In *Geostatistics Troia '92*, Vol. I (ed. A. Soares), 479–492. Kluwer Academic, Amsterdam.
- Vecchia, A. V. (1989). Estimation and model identification for continuous spatial processes. *J. Roy. Statist. Soc. B* **50**, 297–312.
- Veijanen, A. (1990). An estimator for imperfectly observed Markov random fields. Research report, Department of Statistics, Helsinki.
- Walden, A. T. & Guttorp, P. (eds.) (1992). *Statistics in the environmental & earth sciences*. Edward Arnold, London.
- Warnes, J. J. & Ripley, B. D. (1987). Problems with likelihood estimation of covariance functions of spatial processes. *Biometrika* **74**, 640–642.
- Watkins, A. J. & Al-Boutiah, F. H. M. (1990). On maximum likelihood estimation of parameters in incorrectly specified models for covariance for spatial data. *Math. Geol.* **22**, 151–173.
- Yaglom, A. M. (1962). Some classes of random fields in n -dimensional space, related to stationary random processes. *Theory Prob. Appl.* 273–320.

Received May 1992, in final form April 1993

Nils Lid Hjort, Department of Mathematics, University of Oslo, P.B. 1053 Blindern, N-0316 Oslo, Norway.

Henning Omre, Norwegian Institute of Technology, Department of Mathematical Sciences, N-7034 Trondheim, Norway.

DISCUSSION AND COMMENTS

MARIANNE FRISÉN

University of Göteborg

I have a comment about the construction of good predictors. Your measure of the goodness of a prediction, in the section about continuous random surfaces, is the mean squared error, which is denoted by the prediction error, σ_{pe}^2 . You concentrate on replacing the parameters in the model by the maximum likelihood estimates. This procedure does not generally minimize the prediction error. In order to do that you often have to decrease the variance and accept some bias. Then shrinking estimators like Stein-, ridge- and some Bayes methods become interesting. For the somewhat related ARMA-models, some “non-informative” priors have proved valuable for the decrease of the prediction error (Tòmasson, 1986). You did suggest a Bayes method but evaluated it within a Bayesian framework. It would be interesting with an evaluation by the prediction error.

Reference

Tòmasson, H. (1986). *Prediction and estimation in ARMA models*. Almqvist & Wiksell, Stockholm.

FRED GODTLIEBSEN*The Norwegian Institute of Technology, Trondheim*

I would like to congratulate the authors on an impressive paper. I am, in particular, fascinated by the large scale of applications. The authors have shown very good abilities to discover interesting problems, and to realize that these problems can be solved by applying spatial statistics.

My first point will be a few remarks and questions concerning the use of mrf's as prior models for images. Then, I will discuss the use of loss functions for constructing estimators. Finally, I will give some comments on performance evaluation.

Markov random fields*Modelling prior information*

If we compare an image from a typical application, say magnetic resonance (MR) imaging, to a realization from an mrf, it is clear that these two images are very different globally. I believe it will be very difficult to choose an mrf which is able to describe the natural structures, in e.g. a brain, satisfactorily. Because of this difficulty, simple priors are typically preferred. One example is the Geman & McClure (1987) prior model which only involves cliques containing two pixels.

A possible weakness of the Geman–McClure mrf model is that it does not take into account the simultaneous configuration of the region (window) we are looking at. In a recent paper, Godtliebsen & Sebastiani (1992) have constructed a prior model which tries to cope with the above-mentioned weakness. They conclude that, when the image does not have a very complicated structure, better results, compared to e.g. the Geman and McClure approach, can be obtained by applying the new method. I am therefore very happy to see that Hjort and Omre also think that such an approach is well worthwhile. Could the authors comment shortly on their experience and success with methods like those in Tjelmeland & Holden (1993)?

Consistency of priors

In section 3.4.2 the authors claim that it is not serious if a proposed local g -model turns out to be inconsistent with the full model for x . I would appreciate a more precise explanation for this claim than the one given in the present paper.

Loss functions

In image restoration our aim is to estimate x by \hat{x} . Reasonable estimators for x can be found by introducing loss functions. Very frequently, we either use global or local loss functions. But, it seems more natural to use a loss function that has a contextual element, as is also mentioned by the authors. Typically, we could choose a loss function which penalizes the situation that two neighbouring pixels are both misclassified, see Geman (1990). Finding the corresponding estimator explicitly seems very difficult. However, Rue & Frigessi (1992) are able to solve this problem numerically in a very elegant way. It will be very interesting to see whether the choice of loss function really matters for the restoration result.

My belief is that the loss function is a suitable tool, in some cases at least, for constructing estimators but I am not convinced that an intermediate loss function, like the one described

above, will improve the restoration result to a visible degree. What is the opinion of the authors about this?

Performance evaluation

This is a very important, but often neglected, part of the analysis. The evaluation will, however, depend upon the chosen performance criteria. This is illustrated in Godtliebsen (1989) by comparing mean square error and an edge criterion based on differences between neighbours. Moreover, the evaluation will be problem dependent as illustrated in the following examples.

In noise reduction of MR images a good approximation of the true image can be found by making an average of several measurements of exactly the same slice. Objective performance criteria can be obtained by comparing the processed images directly to the average image, see Godtliebsen (1989).

Suppose the aim is to develop an automatic algorithm for drawing the inner contours (blood/muscle interface) in heart images. Then, the method's success can be evaluated by measuring known volumes (phantoms), by checking that contours drawn by the automatic algorithm are consistent with hand-drawn contours, and by checking that the left and the right stroke volumes are found to be equal to normal volunteers.

In reservoir modelling it is more difficult to evaluate the success of the method because the "true" result is not available until several years after our study is performed. My critical question in this application is therefore: how do we know that spatial methods improve the result obtained by traditional methods?

My main question concerning performance evaluation is: how reliable is the evaluation (and thereby our model verification) in situations where we have very few observations (and maybe preferential sampling)?

References

- Geman, D. (1990). *Random fields and inverse problems in imaging*. Lecture Notes in Mathematics, Springer, Berlin.
 Geman, S. & McClure, D. (1987). Statistical methods for tomographic image reconstruction. *Proceedings of the 46th session of the International Statistical Institute, Bull. ISI*, **52**.
 Godtliebsen, F. & Sebastiani, G. (1992). Statistical methods for noisy images with discontinuities. Preprint, The Norwegian Institute of Technology, Trondheim.
 Rue, H. & Frigessi, A. (1992). Personal communication.

JON HELGELAND

Norwegian Computing Centre

Most expositions of spatial statistics tend to focus on linear prediction in the continuous case, and noise filtering in the mosaic case. These techniques are only a part of what the authors appropriately call "predictive statistics". It is most welcome to see a survey with such strong emphasis on the multitude of other applications of this technology, based on the extensive experience of the authors within various domains of application.

Spatial modelling is not very much concerned with estimation theory and hypothesis testing. One reason is of course that parameter estimation uncertainty is of second order in prediction and classification, so it does not make that much difference which estimation method we

choose. Another reason, and perhaps more important, is that the models are in general intractable. In this perspective, the quasi-likelihood methods presented are promising. A simulation study of their performance would be very interesting. Spatial models are often used by people without strong statistical background, meaning that the availability of simple, reliable, and more or less automatic estimation methods is important.

Confirmatory statistics are not quite irrelevant in the spatial setting. An example that springs to mind is the analysis of acid deposition and acidity of Norwegian lakes. The deposition of sulphur and nitrogen oxide can be regarded as spatial explanatory variables. A basic question is whether they have any effect at all on the pH of the lakes. This is an instance of the basic Kriging model of 3.1.1. In this kind of analysis, the spatial model really matters. Strong spatial residual correlations may result in non-significant β s, and a very different conclusion. The problem is well known to time series practitioners.

Robustness is related to model fitting. There are spatial models that in fact display a certain lack of robustness, as witnessed by overshoots and oscillations of interpolated surfaces. These models assume very smooth stochastic fields, and it is not so obvious in practice when we should rely on them, and when we should not.

The Bayesian Kriging approach of section 3.1.3 seems very useful. The examples show how spatial information from several sources can be combined. A possible, important application is the following. In air pollution modelling, it is common practice to use a numerical model to predict the transport of pollutants. Thus, the EMEP data of section 4.4 are predicted by a very complex set of partial differential equations. The air quality experts tend to favour the deterministic, physical model to the probabilistic, agnostic ones. They do, however, appreciate the fact that the models are somewhat imprecise and at the same time, fairly "stiff" and difficult to adjust according to measured data. Various approaches to this "data assimilation" problem can be found in the literature, usually based on some kind of spatio-temporal statistical approach, such as Kalman filtering or Kriging. The present theory suggests a consistent and clear-cut way of doing model adjustments. Assume that the actual concentration is given by

$$Y(x, t) = \alpha + \beta\xi(x, t) + \varepsilon(x, t)$$

where $\xi(x, t)$ denotes the output from the numerical model. Fitting this equation, with or without a prior on α and β , automatically calibrates the numerical model. The technique can also be elaborated, in section 4.4, to incorporate temporal drift or spatial non-stationarity. In view of the relative abundance of observational data, this problem may not be truly Bayesian in nature, but the possibility of deliberately weighting the model information is worth pursuing.

This discussant would like to see more work being done by statisticians on bridging the gap between the very empirical and the very physical modelling approaches. So far, the often sophisticated probabilistic work in this area has been left to others. There is still much work to be done, particularly on model validation and uncertainty. Some of the problems are in fact statistical in nature. In the case of e.g. air pollution transport, it is the probability distribution of pollutant concentration at a particular location, and the way this distribution depends on the various material sources, that matters.

JESPER MØLLER

University of Aarhus

The authors are to be congratulated on a broad overview on spatial statistics including a wide range of applications. Below I would like to discuss some specific aspects on modelling spatial processes and the use of Markov chain Monte Carlo methods.

Hjort and Omre mention some alternative models to mrf's. Especially, in section 3.2.5, they discuss a specific semi-mrf model which may look somewhat complicated. Such models may be considered as being Markov if the local Markov property is extended to include situations where the concept of neighbours is dynamic and not as usual restricted to a fixed relation. Appropriate definitions and a discussion of such "extended" Markov models were presented in Baddeley & Møller (1989) using a general set up for spatial point processes which in fact includes lattice processes as a special case.

In order to show a simple example of a model which is Markov in the extended but not in the "classical" sense, consider a random disc process $\mathbf{x} = \{x_1, \dots, x_n\}$; this can be formalized as a marked point process, cf. Example 4.6 in Baddeley & Møller (1989). Then it may be natural to assume that we can only observe the path-connected components in the union $x_1 \cup \dots \cup x_n$ of the discs; let us call these components for the clusters in \mathbf{x} . Suppose also that the density $f(\cdot)$ for the disc process is strictly positive, and define the conditional intensity $\lambda^*(x, \xi) = f(\{x_1, \dots, x_n\})/f(x)$ of a disc ξ given $\mathbf{x} = \{x_1, \dots, x_n\}$. Then a simple but intuitive appealing assumption might be that $\lambda^*(x, \xi)$ depends only on the number of clusters in \mathbf{x} which intersect the disc ξ . It can be shown that this is equivalent to specifying the density by

$$f(x) \propto \exp \{ \theta_1 n(x) + \theta_2 c(x) \} \quad (1)$$

for some parameter $(\theta_1, \theta_2) \in \mathbb{R}^2$ and where $n(x)$ and $c(x)$ denote the number of discs and clusters in \mathbf{x} , respectively. Further details will be published elsewhere (Møller, 1993), but simulations confirm that (1) provides a flexible class of models for attraction when $\theta_2 < 0$ as well as repulsion when $\theta_2 > 0$.

Hjort and Omre state that pairwise interaction processes provide a rich class of models for point patterns. I agree in the case of modelling repulsion but not in the opposite case: roughly speaking, using a notation as in section 3.3.1, interaction terms $h(\|x_i - x_j\|)$ which aim at modelling attraction should be greater than 1 if x_i and x_j are close (but eventually not too close). Hence realizations may typically look like massive clusters as it is most likely that $h(\|x_i - x_j\|) > 1$ for all pairs $\{x_i, x_j\}$ of points in \mathbf{x} . For example, for specificity and simplicity, consider a Strauss process conditional on the number n of points so that at least its conditional density

$$f_n(x) \propto \gamma^{s_\delta(x)} \quad (2)$$

is well-defined when $\gamma > 1$, see Kelly & Ripley (1976). Here $s_\delta(x)$ denotes the number of pairs $i < j$ with $\|x_i - x_j\| < \delta$ where the parameter $\delta > 0$ is the interaction radius. Then simulations show (i) an extremely strong attraction with the points in mostly one clump if γ is large, while (ii) values of γ close to 1 of course produce realizations which look like a binomial process, whereas (iii) more "interesting" patterns of attraction may only occur for γ in a very narrow subinterval of $(1, \infty)$. See also Geyer & Thompson (1993). Moreover, it does not "help" to include a hard-core parameter $\varepsilon \in (0, \delta)$ so that the conditional density $f_n(x)$ is given by (2) if $s_\varepsilon(x) = 0$ and $f_n(x) = 0$ otherwise: except that the points are now separated by the hard-core distance ε , simulated realizations show similar features as in (i)–(iii).

More interesting models for attraction should be considered. Triple interaction processes provide one possibility. Then attraction may be modelled as described above with pairwise interaction terms greater than 1, but with triple interaction terms smaller than 1 so that massive clustering becomes less likely. This may also ensure that the unconditional density is well-defined. On the other hand, one may feel a need for other ways of modelling attraction which are based on simple but appealing constructions like the derivation of the disc process

(1) as described above. This suggests including interaction terms for maximal cliques so that the density is Markov in the sense of Baddeley & Møller (1989).

Hjort and Omre describe various algorithms for simulations. The systematic updating schemes considered in sections 3.2.3 and 3.2.4 do not generate a reversible Markov chain; in fact the Markov chain generated by certain Metropolis–Hastings algorithms may even be reducible and hence non-ergodic when a cyclic updating scheme is used. Of course reversibility is not always needed, but it may be desirable when using central limit theorems (Kipnes & Varadhan, 1986) or for studying Monte Carlo errors (Geyer, 1992). In order to retain reversibility, either random updating as in the original Metropolis algorithm or systematic schemes like forward–backward scanning (Besag, 1986) can be used.

In particular, Hjort and Omre consider the Gibbs sampler. Despite its popularity the Gibbs sampler is not always the simplest and most efficient single component updating algorithm, cf. e.g. Peskun (1973) and Frigessi *et al.* (1993). Further, in practice Gibbs sampling often involves rejection sampling which in turn can involve a low acceptance rate. For example, consider the “Ripley–Kelly birth-and-death technique” in section 3.3.2, which is yet another example of Gibbs sampling but with random updating. This becomes extremely slow in the case of the conditional Strauss process (2) with γ a bit larger than 1 as uniformly generated proposals are very often rejected when using rejection sampling. The standard Metropolis algorithm is easier to use. However, both algorithms get stuck in one clump of points for a huge number of iterations if γ is large. Geyer & Thompson (1993) describe annealing-like samplers which become more efficient.

Hjort and Omre do not discuss how to simulate spatial point processes with a random number of points. The usual technique is to run a spatial birth–death process (Preston, 1977; Ripley, 1977; Lotwick & Silverman, 1981; Møller, 1989), while the possibility for using a Metropolis–Hastings type algorithm has recently been studied in Geyer & Møller (1994). In that paper we argue that the latter can be much simpler to use and analyze and it is probably also much faster to run. Of course, spatial birth–death processes may still be of interest for other purposes. For instance, Møller & Sørensen (1994) use them to model linear dune fields in the Kalahari Desert.

Markov chain Monte Carlo methods make statistical inference feasible for complex statistical models as used in spatial statistics as well as other areas of statistics. Especially, Monte Carlo maximum likelihood is a useful method for estimating the parameters of the model and investigating the statistical properties of the estimators. This becomes in particular important for spatial models. For example, for most point processes, it is not known whether there is a central limit theorem for the mle as the size of the window containing the observed point pattern is increased; results seem to be available only for processes which are very close to the Poisson processes (Jensen, 1991, 1993). By the way, significance testing based on Monte Carlo likelihood ratio methods has to the best of my knowledge so far not been used in spatial statistics, although this seems to be an obvious idea.

Another application of Markov chain Monte Carlo is to “verify” heuristic justifications. For instance, in most statistical analysis of spatial point patterns, inference is performed conditional on the number of points since one usually believes that this statistic provides little information about the interaction structure. Simulations confirm this for the Strauss process (Geyer & Møller, 1994), but it is not known whether this is true for general processes.

Finally, Hjort and Omre notice that conventional Markov chain Monte Carlo algorithms can be very slow and they call for more efficient algorithms. Geyer & Thompson (1993), Besag & Green (1993), and the accompanying discussions and references deal with this and many other issues. Another source on the more practical aspects is provided by Geyer (1992).

References

- Besag, J. & Green, P. J. (1993). Spatial statistics and Bayesian computation (with discussion). *J. Roy. Statist. Soc. B* **55**, 25–37.
- Frigessi, A., di Stefano, P., Hwang, C.-R. & Shen, S.-J. (1993). Convergence rates of the Gibbs sampler, the Metropolis algorithm and other single-site updating dynamics. *J. Roy. Statist. Soc. B* **55**, 205–220.
- Geyer, C. J. (1992). Practical Markov chain Monte Carlo (with Discussion). *Statist. Sci.* **7**, 437–511.
- Geyer, C. J. & Møller, J. (1994). Simulation procedures and likelihood inference for spatial point processes. *Scand. J. Statist.* to appear.
- Geyer, C. J. & Thompson, E. A. (1993). Annealing Markov chain Monte Carlo with applications to pedigree analysis. Technical report No. 589, School of Statistics, University of Minnesota.
- Jensen, J. L. (1991). A note on asymptotic normality in the thermodynamic limit at low densities. *Adv. Appl. Math.* **12**, 387–399.
- Jensen, J. L. (1993). Asymptotic normality of estimates in spatial point processes. *Scand. J. Statist.* **20**, 97–109.
- Kelly, F. P. & Ripley, B. D. (1976). A note on Strauss' model for clustering. *Biometrika* **63**, 357–360.
- Kipnes, C. & Varadhan, S. R. S. (1986). Central limit theorem for additive functionals of reversible Markov processes and applications to simple exclusions. *Comm. Math. Phys.* **104**, 1–19.
- Møller, J. (1989). On the rate of convergence of spatial birth-and-death processes. *Ann. Inst. Statist. Math.* **41**, 565–581.
- Møller, J. (1993). A note on spatial point processes and Markov chain Monte Carlo. In preparation.
- Møller, J. & Sørensen, M. (1994). Statistical analysis of a spatial birth-and-death process model with a view to modelling linear dune fields. *Scand. J. Statist.* **21**, 1–19.
- Peskun, P. H. (1973). Optimum Monte-Carlo sampling using Markov chains. *Biometrika* **60**, 607–612.
- Preston, C. J. (1977). Spatial birth-and-death processes. *Bull. Int. Statist. Inst.* **46**, 371–391.

EVA B. VEDEL JENSEN

University of Aarhus

I would like to compliment the authors for a very interesting and stimulating paper. My comments are complementary to the important subjects presented in the paper.

The authors do not touch upon the problem that many 2-D images, e.g. microscopical images obtained by either light or electron microscopy, are sections of a 3-D structure. Such images call for stereological analysis. Since most research results in stereology are published in journals not immediately available to statisticians, I will take the opportunity to summarize shortly some of the recent advances in stereology.

Stereology does not aim at 3-D reconstruction of the spatial structure, but uses that random sections contain information about quantitative parameters of the spatial structure in a statistical sense. For instance, in cell biology, the number of cells and their mean volume can be estimated stereologically from observations on pairs of parallel sections, viz. a so-called di-sector, without specific assumptions about the shape of the cells (Sterio, 1984). Other quantitative parameters, which can be stereologically estimated, include length, surface area, curvature and connectivity.

Stereology has close connections to survey sampling theory and analogues exist of finite and infinite population inference and of superpopulation inference (Baddeley, 1993). Systematic sampling is also possible in stereology, e.g. in the form of a set of equally-spaced parallel sections planes. The prediction of the variance of stereological estimators obtained under such systematic sampling, for instance using Matheron's approach (1965), is a very active area in stereology at the moment (Cruz-Orive, 1993).

Classical stereological methods are well-suited for estimating bulk parameters such as the total volume of a cell aggregate but unrealistic assumptions about cell shape have earlier been needed for the estimation of parameters such as the cell number and mean size of cells.

Within the last ten years, new stereological methods of analysing cell populations have been developed which are not shape-dependent. Some of these methods use the possibility of observing on a continuum of parallel thin optical sections, which may be obtained by moving the focus plane in the microscope up and down through the specimen of the structure. Using a tandem scanning reflected light microscope, this can be done in unprepared biological tissue. In fact, this technology of entering the third dimension has released a series of new stereological methods (Cruz-Orive & Weibel, 1990; Jensen & Kieu, 1994). Modern stereological methods have also been applied to 2-D images obtained by computerized tomography (Pache *et al.*, 1993).

Explicit parametric stochastic models are seldom used in stereology. See however Stoyan (1993). One reason is that available models cannot describe the high degree of organization seen in real (especially biological) microscopical structures.

References

- Baddeley, A. J. (1993). Stereology and survey sampling theory. *Bull. Intern. Statist. Inst. Proceedings of the 49th session of ISI*, Florence.
- Cruz-Orive, L. M. (1993). Systematic sampling in stereology. *Bull. Intern. Statist. Inst. Proceedings of the 49th session of ISI*, Florence.
- Cruz-Orive, L. M. & Weibel, E. R. (1990). Recent stereological methods in cell biology: a brief survey. *Amer. J. Physiol.* 258, L148–L156.
- Jensen, E. B. V. & Kieu, K. (1994). Unbiased stereological estimation of d -dimensional volume in R^n from an isotropic random slice through a fixed point. To appear in *Adv. Appl. Prob.*
- Pache, J. C., Roberts, N., Vock, P., Zimmermann, A. & Cruz-Orive, L. M. (1993) Vertical LM sectioning and parallel CT scanning designs for stereology: application to human lung. *J. Microsc.* 170, 9–24.
- Sterio, D. C. (1984). The unbiased estimation of number and sizes of arbitrary particles using the disector. *J. Microsc.* 134, 127–136.
- Stoyan, D. (1993). Stereological analysis of particle systems by planar sections. *Bull. Intern. Statist. Inst. Proceedings of the 49th session of ISI*, Florence.

MATS RUDEMO and HENRIK STRYHN

Agricultural University, Copenhagen

Nils Lid Hjort and Henning Omre have given a highly useful survey of statistical methods in spatial problems with an admirable interplay between practical examples and theory. Typical issues in spatial applications are large amounts of data, often with a high degree of structure, which is difficult to model adequately. This makes spatial statistics an ideal proving ground for development of estimation, prediction, classification and simulation methods and algorithms that should be useful also in other application areas. In the present discussion we will compare genetic models with spatial models, and then we will, as a complement to the models presented by Hjort and Omre, briefly review some statistical models aimed at image segmentation problems.

Let us first describe an application in population genetics where the problems are very similar to some of those discussed in Hjort and Omre's paper. Using notation as in their equation (3.1), we let

$$Z(x) = m(x, \beta) + e(x), \quad x \in D,$$

denote a vector of continuous (polygenic) traits measured on an animal x in a finite population D . Typically $m(x, \beta)$ includes herd–year–season effects while the covariance

function of ε is determined by genetic relations among the individuals. Then the Kriging equations correspond to the BLUP (best linear unbiased prediction) method for estimating breeding values, see, e.g., Robinson (1991).

It may be noted that from classical genetic theory it follows that $Z(x)$, $x \in D$, is a Markov random field, where the neighbours of an individual consist of parents, children and partner(s). This is one of the rather few cases where there is a simple theoretical motivation for the Markov property on a non-linear lattice—some other examples come from physics. With algorithms presently available it is possible to compute breeding values for continuous traits even for very large populations (several million observations). However, to compute covariances or, in particular, to estimate breeding values for qualitative traits controlled by major (or minor) genes, only small populations can be handled. Here it seems that some of the methods described in Hjort and Omre's paper could be highly useful. For instance, we think it would be very interesting to try the quasi-likelihood methods of section 3.1.5 and the simulation techniques described in sections 3.2.3–4 in genetic applications. Conversely the recursive updating techniques developed in genetic contexts (Andersen, 1985) and for expert systems (Lauritzen & Spiegelhalter, 1988) could be quite useful also in geostatistics.

For estimation of mixed model variance components the REML (restricted maximum likelihood) method is closely associated with BLUP as a prediction method, cf. the comment by Terry Speed in the discussion of Robinson (1991). Partly due to the recent availability of suitable programs, the REML method is now much used for parameter estimation in genetic applications. It seems as REML could be a useful (minor) modification in section 3.1.5 of the maximum likelihood method minimizing formula (3.10). In fact, the results of Stein (1990b) indicate that, at least for correctly specified smoothing spline models, REML is an attractive method (better than generalized cross-validation).

Our next comment sets off from the need stated by Hjort and Omre for alternatives to the Markov random field (mrf) approach to image segmentation problems. There has been some scepticism raised in the statistical community against segmentation as a preprocessing step in image analysis, see Geman & Gidas (1991); however, in the medical applications mentioned by Hjort and Omre the segmented objects are of intrinsic interest, and the procedure seems well justified. But why do the mrf models not always give satisfactory solutions? A plain answer is that although observations are discrete, usually situated at an equidistant lattice, the problem is not. The boundaries to be estimated often have continuous scale-invariant properties, for instance related to shape or smoothness. Some control of such features can be obtained in mrf models via the edge process by penalization. However, continuous prior information can be utilized more directly using the Bayesian paradigm; examples are the polygonal mosaic models introduced by Arak & Surgailis (section 3.2.5) as well as models built by deformable templates depicting particular shapes, e.g. of hands, potatoes or nematodes. An abstract and very flexible class of models of this type has been developed by Ulf Grenander and co-workers; for the first of the mentioned applications, see Grenander *et al.* (1991).

Smoothness of and discontinuity at the boundaries have been cornerstones in a recent development of what might be termed spatial change point models. They generalize the classical change point problem to two- or multi-dimensional "time" simply by dividing the scene into regions with statistically coherent observations, i.e. obeying a common distributional law. Distributions within regions are quite arbitrary; present examples include independent and stationary or regression-type observations as well as random fields subject to mixing conditions. For the boundaries some regularity conditions must be imposed, either explicitly in a non-parametric set-based formulation or implicitly via direct boundary parametrization. Within set-ups combining some of these features theoretical results like consistency, (optimal)

rate of convergence and asymptotic distribution of boundary estimators have been obtained. Note that regular lattice observations restrict the obtainable convergence rates, a fact that has motivated models involving some randomness in the spatial locations or boundaries tied to the observation lattice. The most complete results are given in Korostelev & Tsybakov (1993), integrating recent developments on asymptotic optimality. Other treatments and further references can be found in the conference proceedings edited by Carlstein *et al.* (1993).

As our final point, a less direct but sometimes advantageous approach to discrete spatial problems is at first to study the continuous observation analogues, for which in general—as in the unilateral case—more clear-cut solutions are feasible. For instance, Zhang & Stein (1993) analyze m th order intrinsic random function models (equivalent to thin plate smoothing) to obtain kernel Kriging predictors that are shown to provide good approximations for observations on a square lattice. For spatial change point models a continuous Gaussian white noise treatment has been initiated by Hasminskii & Lebedev (1990) who give asymptotic distributions for some Bayesian and maximum likelihood boundary estimators.

References

- Andersen, S. (1985). Estimation of breeding values by multitrait direct updating. *Z. Tierzücht. Züchts-biol.* **102**, 81–90.
- Carlstein, E., Müller H.-G. & Siegmund, D. (eds) (1993). *Change-point problems*. (Preliminary title) IMS Lecture Notes—Monograph Series. Hayward, California.
- Geman, D. & Gidas, B. (1991). Image analysis and computer vision. In *Spatial statistics and digital image analysis* (ed. Board on Mathematical Sciences, National Research Council), 9–36. National Academy Press, Washington, DC.
- Grenander, U., Chow, Y. & Keenan, D. M. (1991). *Hands: a pattern theoretic study of biological shapes*. Springer, New York.
- Hasminskii, R. Z. & Lebedev, V. S. (1990). On the properties of parametric estimators for areas of a discontinuous image. *Probl. Control Information Theory* **19**, 375–385.
- Korostelev, A. P. & Tsybakov, A. B. (1993). *Minimax theory of image reconstruction*. Lecture Notes in Statistics, **82**, Springer, Berlin.
- Lauritzen, S. L. & Spiegelhalter, D. J. (1988). Local computations with probabilities on graphical structures and their application to expert systems (with discussion). *J. Roy. Statist. Soc. B* **50**, 157–224.
- Robinson, G. K. (1991). That BLUP is a good thing: the estimation of random effects (with Discussion). *Statist. Sci.* **6**, 15–51.
- Zhang, B. & Stein, M. (1993). Kernel approximations for universal kriging predictors. *J. Multivariate Anal.* **44**, 286–313.

Marianne Frisén, University of Göteborg, Department of Statistics, Viktoriagatan 13, S-41125 Göteborg, Sweden

Fred Godtliebsen, Universitetet i Trondheim, Norges Tekniske Høgskole, Institutt for Matematiske Fag, N-7034 Trondheim, Norway

Jon Helgeland, Norwegian Computing Centre, P.B. 114, Blindern, 0314 Oslo, Norway

Jesper Møller, Department of Theoretical Statistics, Institute of Mathematics, University of Aarhus, Ny Munkegade, DK-8000 Aarhus C, Denmark

Eva B. Vedel Jensen, Department of Theoretical Statistics, Institute of Mathematics, University of Aarhus, Ny Munkegade, DK-8000 Aarhus C, Denmark

Mats Rudemo & Henrik Stryhn, The Royal Veterinary & Agricultural University, Department of Statistics, Thorvaldsenvej 40, 1871 Copenhagen, Denmark

REJOINDER TO THE DISCUSSION

NILS LID HJORT and HENNING OMRE

Marianne Frisén

As Frisén mentions, the objective is to construct a good predictor. Generally, optimality is reached by using the conditional distribution. A natural choice is to use conditional expectation with conditional variance as best prediction and measure of prediction uncertainty. The model used by us is within the class of multinormal models, and the model parameters are assumed known. Hence the conditional distribution is Gaussian with parameters uniquely defined by the conditioning values and the model parameters. In particular, the best prediction will be a linear combination of the observed values. This prediction will be optimal under a large variety of loss criteria due to the unimodality and symmetry of the conditional distribution.

Frisén makes some interesting comments about prediction in ARMA-models. These models are somewhat different from the ones we discuss in section 3.1, however. The autoregressive moving average models are defined by the number of non-zero terms in the predictor, and one of the aims is to construct optimal predictors under such constraints. In our case the optimal predictor is a consequence of the underlying multi-Gaussian model. If a wider scenario is envisaged, where selection of the model is part of the process, say selection of the number and the form of the f_j functions in equation (3.1), then other predictors can perform better.

Fred Godtliebsen

Godtliebsen mentions the shortcoming of Markov random fields when it comes to actually describing the often quite complicated physical images for which they are nevertheless used; simulate from a simple mrf and you won't see a human brain. We agree with this point, and have not claimed that all the prior distributions we have employed in various applications have been completely realistic. See also Rudemo and Stryhn's comments. One may speculate that the reason why using quite simplistic priors of the mrf type are useful in image analysis is that they after all often manage to capture the local homogeneity structure of the image; what matters for the final analysis is the posterior distribution, which will be realistic with sufficient data. Godtliebsen further states that "because of this difficulty, simple priors are typically preferred". Our view is that simple priors are used since people can understand them and handle them, which is a good start. When priors are acknowledged to be imperfect in certain ways that should be an impetus to work out better alternative models, and attempts have indeed been made to construct more realistic priors for images, cf. our sections 3.2.4–6.

Formulae (3.24), (3.25) and (3.28) represent examples of simple methods for contextual classification that exploit neighbourhood information for the four nearest neighbours. At the end of section 3.4.2 we make a case for perhaps pragmatically extending these to fuller neighbourhoods, even if it should lead to formulae that are not fully consistent with exact models, and Godtliebsen queries this. An example is the eight-neighbour generalisation of (3.24), taking the form

$$P_i\{k \mid y_{N(i)}\} = \text{const } \pi(k)f(y_i \mid k) \prod_{j=1}^4 T_k(y_{i,j}) \prod_{j=5}^8 U_k(y_{i,j}). \quad (1)$$

Here the four immediate neighbours $y_{i,1}, \dots, y_{i,4}$ and the four diagonal neighbours $y_{i,5}, \dots, y_{i,8}$ together make up $N(i)$, and

$$T_k(y) = \sum_{m=1}^K p(m | k) f_m(y) \quad \text{while} \quad U_k(y) = \sum_{m=1}^K q(m | k) f_m(y). \quad (2)$$

Here $p(m | k)$ is the probability that the immediate neighbour of a pixel from class k should be from class m , and correspondingly $q(m | k)$ for a diagonal neighbour. These latter transition probabilities could be modelled and estimated separately, or one could use $q(m | k) = \sum_{l=1}^K p(l | k) p(m | l)$, suggested by a conditioning argument. One would expect an alternative method based on the appropriate generalisation of (3.23) to be even better. This must involve full specification of configuration probabilities $g(a_1, \dots, a_8 | k)$, however. Our point is that this would typically be a hard task, both modelling-wise and computationally, while the pragmatic approximation (1) with (2) defines a simpler and perfectly valid contextual classification rule, often having smaller pixel-wise error rate than the four-neighbour version (3.24).

We have illustrated this point by computing the exact classification error rates for the four-neighbour method and the eight-neighbour method, in situations with two equally likely multi-normal populations on the scene. The two classes have the same covariance matrix, and Nature is assumed to distribute the two classes over the scene according to a certain Poisson field model described in Owen (1984) and Hjort (1985b). Figure 11 displays the exact error rates for the four-neighbour and eight-neighbour methods as a function of the Mahalanobis distance δ between the two normal populations. Also included is the error rate $\Phi(-\frac{1}{2}\delta)$ of the traditional non-contextual discriminant method. The figure shows these curves for two situations with different degrees of spatial homogeneity; the percentage of homogeneous 3×3

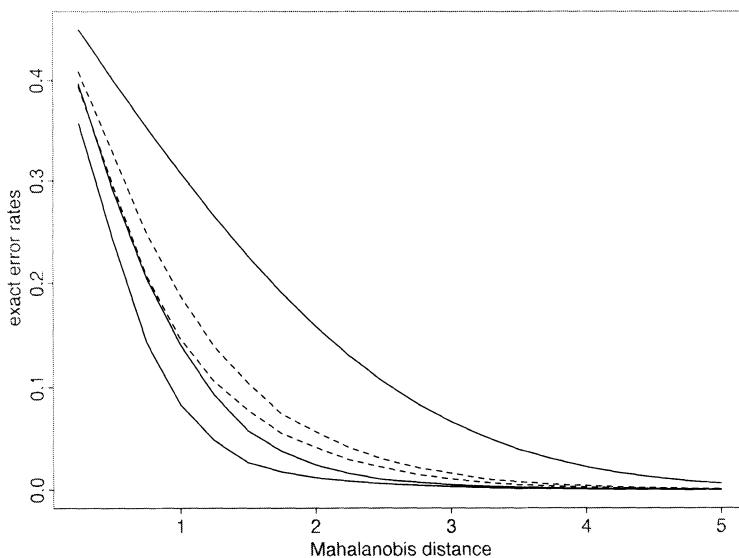


Fig. 11. Exact error rates for the four- and eight-neighbour multiplicative classification rules, as a function of Mahalanobis distance between the two multinormal populations, in two situations: a case of strong spatial continuity for the classes, with 90% of 3×3 pixel boxes being homogeneous (the two fully drawn curves), and a case of weaker spatial continuity, with 55% of 3×3 pixel boxes being homogeneous (the two dotted lines). The top curve is the exact error rate for the ordinary non-contextual discriminant analysis method.

pixel-boxes is respectively 90% and 55%. Note that the spatial image model has been used here only to compute exact error rates and not to construct the classification methods themselves. The solid improvements of the four-neighbour method over the non-contextual one, and of the eight-neighbour method over the four-neighbour method, illustrate the usefulness of such pragmatically constructed spatial classifiers.

There are two intimately connected components of a Bayesian decision problem, the prior and the loss function, and perhaps research in traditional Bayesian robustness has concentrated too much on robustness with respect to the prior and too little on changes in the loss function. This is a pertinent remark in Bayesian image analysis, where there is a dramatic difference between the two ways loss functions (3.21) and (3.26) view the quality of a classification, and we agree with the discussant that alternative loss functions should be tried out. Rue (1993) shows how optimal Bayes images can be arrived at for reasonably general loss functions, in the two-class case and with a simple mrf as prior. This needs to be generalised to the multi-class case and tried out in a broader range of case studies, also with respect to the prior mrf employed. See also sections 2D and E in Hjort & Mohn (1987), where procedures stemming from an intermediate loss function, caring only about correctly classified 2×2 blocks, are analysed. Derin *et al.* (1984), on the basis of extensive simulations in the one-dimensional Markov chain case, report that solutions based on the pixel-wise criterion performed "just as well, if not better", than the simultaneous maximum *a posteriori* probability solution. Preston (1971) similarly noted in an early paper (also on one-dimensional problems) that the global criterion performed well under the local loss function too. That the global "all or nothing" loss function is sometimes too harsh is shown by some of the examples in Greig *et al.* (1989).

Godtliebsen's questions concerning performance evaluation are important. In reservoir modelling there are obvious problems when the production is completed only several years later. And even at that stage the "truth" about the reservoir geology will not be known. There is hope, however, since during the development phase up to one hundred wells may be drilled and the geology observed. Moreover, production is monitored continuously from the start of production, providing growing insight into the geology. Lastly, the geological formations are repeated in many North Sea reservoirs, allowing for geological inference between the reservoir under study and previously produced ones. It is also worth noting that predicting production is not the only purpose of reservoir evaluation. Often the objective is to evaluate the relative sensitivity with respect to different geological variables. This is simpler than prediction of volumes since only the ordering is called for.

Jon Helgeland

We are generally in agreement with Helgeland's comments. He reminds us that there are important problems where an old-fashioned hypothesis test is called for, and where fuller reliance on the actual model used becomes crucial. Perhaps we involuntarily downplayed this "confirmatory statistics" aspect too much through our choice of examples.

Eva B. Vedel Jensen

Our paper has perhaps been a stereological tour of spatial statistics, by necessity examining only lower-dimensional slices of the richer world of all spatial statistics. Jensen does us a service by bringing in another such slice, that of statistical stereology. This field is also moving with the pace of new technological tools, as witnessed by the work reported on in her list of recent references. Chapter 9 in Ripley (1981) is still a good account of the basic theory.

In some of the applications of stereology the shape of the objects, perhaps both the original three-dimensional and its two-dimensional projections, is of relevance. Then it becomes important to describe such shapes and how they vary between objects. Some aspects of such a statistical shape theory are in Kendall (1989). Other models and methods for geometrical probability that could be fruitful in such studies are surveyed in Solomon (1979).

Jesper Møller

Møller raises many points highly relevant for our paper. The field of simulation algorithms for complex stochastic models is in rapid growth, with participation from both statistics and computer science. Møller's supplementary list of mostly quite recent references testifies to this, and will be of interest to our readers. Several of Møller's comments pertain to details of recent improvements of simulation methods, and it will be outside the scope of our rejoinder to comment further on all of these ideas. We use the opportunity to mention a valuable review paper by Neal (1993) which highlights recent advances in the similar and somewhat parallel area of Markov chain Monte Carlo methods. Recently simulation approaches based on the annealing technique have been applied in reservoir modelling too, to cope with complex conditioning, see Deutsch & Journel (1993) and Hegstad *et al.* (1993).

Møller comments rightly that the semi-Markov random field we describe can be viewed as simply a Markov random field if the neighbourhoods are dynamically defined. The objective is to define a mosaic model with a parameterisation closely related to expert intuition. This entails specifying the type of "tiles" in the mosaic by their sizes and neighbourhood relations. The idea of using an extended Markov model in a spatial point process setting, in the sense of Baddeley & Møller (1989), is interesting. For us it is not obvious how to create tiles of different types and size characteristics, however. We have tried with a marked point process, assigning the tile labels and then let the influence of each marked point grow out from the point with different velocities according to the mark. This creates strange, often non-convex, shapes of the tiles, which for our application is unwanted. Moreover, anisotropies beyond simple rescaling of the axis system is frequently required and the spatial point process approach does not simply represent this.

Møller's comment on the crucial difference between repulsive and attractive forces in point processes is pertinent. The repulsive-force models are relatively tractable, while too simple formulations of attractive-force models will result in unwanted characteristics of the realisations. We have some experience with these problems, but we have not tested out the multiple interaction approach. For us it has seemed more natural to seek a solution among doubly stochastic spatial processes. Spatial generalisations of the Cox processes go back to Matérn (1971). A version described in Grandell (1976) may have a potential for modelling clustered point patterns. Such ideas are developed and results reported on in Tjelmeland & Omre (1994).

Parameter estimation in spatial models is far from trivial and very few analytical results are available. Monte Carlo-based approaches are of course tempting, but the extremely high dimensionality of the problems may even make this approach unfeasible from a processing point of view. We agree that Monte Carlo likelihood ratio methods are worth working with, at least on problems of moderate size.

Mats Rudemo and Henrik Stryhn

It is nice to have pointed out application areas such as population genetics where the Markov property follows from basic theoretical laws rather than just being "assumed" on the

grounds of mathematical-statistical convenience. We would be delighted if some of the methods we describe could prove useful in such areas.

There is a need for models that are less scale- and lattice-dependent than the mrf ones for use in image analysis, cf. our comments in section 3.2.5 and elsewhere. We are grateful for the supplementary information and insight the discussants provide in this direction. Their recent paper (Rudemo & Stryhn, 1993) is one such recent model formulation and solution to a spatial change point problem. There is also much current interest in new ways to represent and handle knowledge of complex patterns and systems. The discussants mention Grenander's work; a recent reference is the discussion paper Grenander & Miller (1994). Of potential benefit in this area is also the rapidly growing field of statistical neural networks, see e.g. Ripley (1993).

References

- Derin, H., Elliott, H., Christi, R. & Geman, D. (1984). Bayes smoothing algorithms for segmentation of binary images modelled by Markov random fields. *IEEE Trans. Pattern Anal. Machine Intell.* **6**, 707–720.
- Deutsch, C. & Journel, A. G. (1993). The application of simulated annealing to stochastic reservoir modelling. *Soc. Petroleum Engrs, Adv. Techniques Ser.* to appear.
- Grandell, J. (1976). *Doubly stochastic Poisson processes*. Lecture Notes in Mathematics. Springer, Berlin.
- Grenander, U. & Miller, M. I. (1994). Representations of knowledge in complex systems (with discussion contributions). *J. Roy. Statist. Soc. B* **56**, to appear.
- Hegstad, B. K., Omre, H. & Tjelmeland, H. (1993). Stochastic simulation and conditioning by annealing in reservoir description. *Proceedings of the workshop on geostatistical simulation* (ed. M. Armstrong). Fontainebleau, France.
- Kendall, D. G. (1989). A survey of the statistical theory of shape [with discussion]. *Statist. Sci.* **4**, 87–120.
- Matérn, B. (1971). Doubly stochastic Poisson processes in the plane. *Statistical Ecology*, I (eds G. P. Patil, E. C. Pielou & W. E. Waters), 195–213. Pennsylvania State University Press, College Town.
- Neal, R. M. (1993). Probabilistic inference using Markov chain Monte Carlo methods. Technical Report, Department of Computer Science, University of Toronto.
- Preston, P. F. (1971). An empirical Bayes problem with a Markovian parameter. *Biometrika* **58**, 535–543.
- Ripley, B. D. (1993). Statistical aspects of neural networks. In *Networks and chaos—statistical and probabilistic aspects* (eds. E. Barndorff-Nielsen, D. R. Cox, J. L. Jensen & W. S. Kendall). Chapman & Hall, London.
- Rudemo, M. & Stryhn, H. (1994). Approximating the distribution of maximum likelihood contour estimators in two-region images. *Scand. J. Statistics*, **21**, 41–55.
- Rue, H. (1993). New loss functions in Bayesian imaging. Part of Dr. Ing. thesis, Norwegian Institute of Technology, University of Trondheim.
- Solomon, H. (1979). *Geometric probability*. SIAM Publications, Philadelphia.
- Tjelmeland, H. & Omre, H. (1994). Simulation algorithms for complex spatial models with conditioning observations. Technical report, Norwegian Institute of Technology, Trondheim.