

Topics in the Statistical Analysis of Remotely Sensed Data

Nils L. Hjort and Erik Mohn

Norwegian Computing Centre, Oslo, Norway

SUMMARY. Satellites and other aircraft are equipped with multi-spectral scanners that can measure spectral reflectance in several separated bands. For each surface element on the ground, therefore, a vector of measurements is available, and information about the properties of the surface, or ground cover, can be inferred from the spectral vectors. The present paper reviews some recently developed spatial models, for the behaviour of ground cover types (classes) over a scene, and for the behaviour of spectral vectors given the scene. These models lead to statistical procedures for classification of picture elements into classes and for prediction of ground variables that are correlated with spectral data. Problems regarding the construction and analysis of such models are discussed.

1. Introduction

Every point on Earth's surface is constantly reflecting and emitting electromagnetic radiation. The intensities vary for the different wavelengths of the electromagnetic spectrum. This spectral distribution, or spectral signature, depends upon several factors, of which the most important ones are surface conditions, the type of land cover, the temperature, the biological activity, and the angle of the incoming radiation. Satellites and other aircraft are equipped with a multispectral scanner, and are able to measure the intensity in several separated narrow bands of the spectrum. The ambitious and difficult task for the analysers of these masses of data is to extract useful information about the surface properties, say classifying a region of land into types of land cover.

This amazingly involved technological process actually works, as witnessed by the enormous and still increasing amount of such data that is now being collected for an increasing range of purposes. This is in spite of severe difficulties with the gathering and preprocessing of remote sensing data, being plagued with distortions stemming from atmospheric absorption, bad weather, pollution, non-perfect spatial resolution, blurring, geometrical errors coming from non-perfect scanner speed, etc.

The data collected in this way are highly structured, consisting typically of a low-dimensional vector of measurements sitting at each of the sites on a rectangular lattice, and many of the tasks to be performed are easily formulated in traditional statistical terms: classification into types of cover, recognition of patterns, detecting change, estimating global and local land parameters, prediction, building models for simulation. Thus in the

processing of such image data statisticians should play an important role. But generally, they are not; as in other technology-driven disciplines we have let engineers and computer scientists take the lead. Only in the last few years have statisticians involved themselves in the many interesting problems of statistical image processing.

This paper discusses only some of the problem areas associated with remotely sensed data, and intends to illustrate how statistical thinking and methodology can serve to find solutions. Section 2 reviews approaches to one of the more prominent problems, that of classifying each picture element (pixel) into one of a finite number of classes. Central in several approaches is the problem of constructing realistic but feasible statistical models for the behaviour of classes over the scene and for the data vectors. The key ingredient is that of spatial context; the intuitive geographical law that “everything is correlated” (and nearby phenomena more so than distant phenomena) must be exploited. Section 3 considers the task of estimating a land parameter of interest for each pixel. Most procedures are again derived from appropriate models whose parameters must be estimated, from training data (i.e. X -measurements for regions of known class) or from unclassified pixel data. Parameter estimation is the topic of Section 4. Finally Section 5 briefly mentions some possible extensions and amendments, and offers remarks on future developments.

References to general issues in remote sensing are Schowengerdt (1983) and Holz (1985). Some recent papers concerned with statistical issues in image analysis (not necessarily remote sensing), usually concentrating on the spatial classification problem, are: Switzer (1980, 1983), Kittler and Föglein (1984), Hjort and Mohn (1984), Geman and Geman (1984), the project report Sæbø et al. (1985), Ripley (1986), Besag (1986), Kay and Titterton (1986), and Green, Berman, Switzer, and Craig (1987). Statisticians are encouraged to seek out some of the many interesting and accessible papers on the subject in journals like *Pattern Recognition*, *Pattern Recognition Letters*, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *IEEE Transactions on Geosciences and Remote Sensing*, *Remote Sensing of Environment*, and *Image and Vision Computing*.

2. Classification

Consider a scene of pixels numbered $i = 1, \dots, N$, with a d -dimensional vector X_i of measurements associated with pixel i . X_i typically consists of the spectral signatures for different bands, but is sometimes augmented with further variables, perhaps related to topography and texture. K types or classes of land cover are defined on the scene, numbered $k = 1, \dots, K$, with c_i the true class label for pixel no. i . X -vectors coming from class k follow a statistical distribution with density $f_k(x)$, usually taken to be the multinormal or Gaussian one. The classification task is to estimate the underlying classes $c_1, \dots, c_N$ on the basis of X 's, possibly after an initial training phase where the class densities are estimated.

Some comments are in order here. First, the X -measurements are not really continuous; the spectral reflectance is usually recorded as a number between 0 and 255. The scale is however fine enough to make the conceptual leap to continuous data acceptable. It also tends to be the case that histograms of X -data from a homogeneous class are well described by the Gaussian density. A more critical assumption, perhaps, is that each pixel

belongs to only one well-defined class; see some comments about mixed pixels in Section 5A. With increased spatial resolution some of the problems associated with these “mixed pixels” become less important, since the relative frequency of non-homogeneous pixels will decrease. The French SPOT satellite has for example a quite impressive resolution corresponding to $20\text{m} \times 20\text{m}$ surface elements, compared to about $100\text{m} \times 100\text{m}$ for the early LANDSATs.

Since obviously some pixels may not belong to any of the predefined classes, we include this possibility by labelling the category of such outliers OUT. These outliers form a potentially very heterogenous class, and it is usually safest to detect them separately; see Hjort (1985a).

2A. The ordinary, noncontextual approach

Standard discriminant analysis of course provides a solution to the classification problem. If $\pi(k)$ is the prior probability of class k , then assigning label k to pixel i , where k maximises $\pi(k)f_k(X_i)$, gives lowest long-term error rate among the rules based on data from the pixel alone. This noncontextual rule is ordinary linear discriminant analysis if the class densities are normal with common covariance matrix, and becomes quadratic discriminant analysis when the covariance matrices are different. These rules are usually included in commercial systems for remotely sensed data. If class separation is good enough, then error rates are low, and the job is done.

2B. 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 $\hat{c}_i$ to pixel i , whose true label is c_i , is of the type

$$L(c_i, \hat{c}_i) = \begin{cases} 0 & \text{if } \hat{c}_i = c_i \text{ (correct decision);} \\ 1 & \text{if } \hat{c}_i \neq c_i \text{ and } \hat{c}_i \leq K \text{ (wrong decision);} \\ t & \text{if } \hat{c}_i = D \text{ (being in doubt).} \end{cases} \quad (1)$$

Thus the possibility of being in doubt and state nothing about c_i is reserved, having in mind, for example, pixels made up of a mixture of points from more than one class. t is a threshold between 0 and 1. Now, if the total loss from a classification, or 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

$$\hat{c}_i = \begin{cases} k & \text{if } k \text{ maximises } P_i\{k|\text{data}\}, \\ & \text{and this maximum exceeds } 1 - t; \\ D & \text{if } P_i\{k|\text{data}\} \leq 1 - t \text{ for each } k. \end{cases} \quad (2)$$

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

In practice one has to limit oneself to a small subset $\text{data}(i)$, containing at least X_i , of all the data. If $\text{data}(i)$ is chosen to consist of X_i and its four immediate neighbours, for example, then the natural approximation to (2) is the rule that maximises

$$\begin{aligned} P_i\{k|\text{data}(i)\} &= \text{const. } \pi(k) f(\text{data}(i)|c_i = k) \\ &= \text{const. } \pi(k) \sum_{a,b,c,d} g(a,b,c,d|k) h(\text{data}(i)|k,a,b,c,d). \end{aligned} \quad (3)$$

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 distribution of classes over the scene. $h(\text{data}(i)|k, a, b, c, d)$ is the density, in dimension $5d$, for the five vectors in question, given that the underlying classes are k, a, b, c, d . The summation in (3) is, in general, over all K^4 configurations. The rule based on (3) is an approximation to (2), 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 $\text{data}(i)$ for the i -th decision.

In this way we are naturally led to consider the class labels $c = (c_1, \dots, c_N)$ as outcomes from a stochastic process $C = (C_1, \dots, C_N)$, governed by some probability model. Similarly $X = (X_1, \dots, X_N)$ is naturally viewed as a stochastic process on the lattice of pixels. For each specification of global, simultaneous models for these processes (for C and for X given C) 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 cross” 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 C , say, we are allowed to view it merely as a convenient approximation to the complex schemes Nature employs when she distributes the classes over the land.

Another typical feature in these problems is also illustrated in formula (3): the models we use must not only be realistic and fittable, but also feasible in the sense of not needing to much cpu-time. A SPOT 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) is the following: suppose that X -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, see Section 4). (The Markov mesh model for classes on a lattice studied by Pickard (1977) has in fact this multiplicative property.) Then (3) simplifies to

$$P_i\{k|\text{data}(i)\} = \text{const. } \pi(k) f_k(X_i) T_k(X_{i1}) \cdot \dots \cdot T_k(X_{i4}), \quad (4)$$

where the X_{ij} ’s stem from the four neighbour pixels and $T_k(x) = \sum_m p(m|k) f_m(x)$. This produces the classification rule reached by Welch and Salter (1971), Haslett (1985), Hjort and Mohn (1984), and others, from somewhat different points of departure.

Hjort and Mohn (1985) obtain a natural generalisation of this rule to the case where spatial autocorrelation between X ’s is allowed for. Specifically,

$$P_i\{k|\text{data}(i)\} = \text{const. } \pi(k) f_k(X_i) U_k(X_i, X_{i1}) \cdot \dots \cdot U_k(X_i, X_{i4}), \quad (5)$$

where the U -functions are appropriate generalisations of the T -functions appearing in (4). They also provide evidence from SPOT data that such autocorrelation is prominently

present with high resolution satellite data and ought to be taken into account. Hjort and Mohn (1984) and Sæbø et al. (1985) consider other variations on theme (3) as well.

The reasoning above applies equally well to larger neighbourhoods than the cross, but exact expressions based on the appropriate generalisation of (3) quickly become long and unfeasible. Again we feel that the statistician should not be afraid of constructing convenient approximations, even if it should lead him outside the safe ground of exact models. For the 3×3 pixel box with eight neighbours, for example, we may use a formula similar to (4), with four more terms entering the product, involving transition probabilities for diagonal neighbours, say $q(\cdot|k)$. ($q(\cdot|k)$ can be expressed in terms of $p(\cdot|k)$; indeed $q(a|k) = \frac{1}{\pi(k)} \sum_l \pi(l) p(k|l)p(a|l)$.) This produces a bona fide classification algorithm with good error rate properties, although it, in fact, cannot be deduced from a bona fide global model for the classes: there is no global distribution for the C process that gives $g(a, b, c, d, a', b', c', d'|k)$ equal to $p(a|k) p(b|k) p(c|k) p(d|k) q(a'|k) q(b'|k) q(c'|k) q(d'|k)$ (in which a', b', c', d' are the diagonal neighbour labels).

*** draw & insert Nils-figure 1 here!!!

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 (1985b) studies one particular eight-neighbour method of the type (3), by an appropriate generalisation of a four-neighbour method due to Owen (1984). Exact formulae for error rates are impossible to obtain, but they can be expressed via probabilities for events involving nine or less (univariate, even if the X 's are multivariate) independent normal variates, 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.) Figure 1 depicts the type of pixelwise error rates that results from respectively pixel cross and pixel box neighbourhood.

2C. Markov random field approaches

The loss function (1) 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

$$L(c, \hat{c}) = \begin{cases} 0 & \text{if every pixel is correctly classified;} \\ 1 & \text{if one or more pixels are misclassified.} \end{cases} \quad (6)$$

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

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

This requires first of all full model specifications for C and X (as opposed to only “local models”). Secondly, it may appear practically impossible to find this maximum a posteriori scene, because of the enormous number K^N of different possibilities to search through. But recent ideas from numerical-probabilistical optimisation make it possible for both Geman and Geman (1984) and Besag (1986), in two important papers, to give satisfactory solutions, for a broad family of prior distributions $p(c)$.

The key ingredient is the model structure for C (and for X given C in cases where the conditional independence assumption cannot be trusted). To define and discuss Markov random fields (MRFs), decide first on a neighbourhood system; one might for example count each of the eight pixels surrounding the centre pixel in a 3×3 box as its neighbours. Say that C is a MRF if the distribution of C_i given the rest of C only depends upon C_j for the neighbours j of i , that is,

$$p_i(c_i | c_{S-i}) = p_i(c_i | c_{\partial i}) \quad (7)$$

in a convenient notation: c_A denotes the collection of c_i ’s with i in the set A , ∂i is used for the neighbour set for pixel i , and S is the whole scene. In particular the neighbourhood system defines the range of possible models.

The remarkable Hammersley-Clifford theorem identifies the MRFs with the so-called Gibbs distributions. To explain this result we need the concept of a *clique*. This is a set of sites such that each pair within the set are neighbours of each other (this requires the neighbour relation to be symmetric). Under a positivity condition it holds that the MRF above must have probability distribution

$$p(c) = \text{const.} \exp \left[\sum_i \alpha_i(c_i) + \sum_Q V_Q(c_Q) \right], \quad (8)$$

where the second sum extends over all possible cliques. The exponent (or sometimes minus the exponent) is called the *energy function* and the V_Q contributions are called *potentials*, with vocabulary brought in from statistical physics. Besag (1974) is a good reference.

For illustrational purposes we discuss mainly Besag’s prime choice corresponding to $V_{i,j}(c_i, c_j) = \beta I\{c_i = c_j\}$ for neighbours i, j (then $\{i, j\}$ is a clique) and $V_Q = 0$ for cliques with more than two elements. Let furthermore $\alpha_i(k) = \alpha(k)$ be independent of pixel location. (Again, this is mostly for expositional convenience; in many applications it will pay to vary the $\alpha(k)$ ’s. $\alpha(k)$ is tied to the expected proportion of class k pixels, over the

scene. Varying the prior probabilities $\pi(k)$ over the scene is the idea of Switzer, Kowalik, and Lyon (1982), and can be made operational in this MRF setting too.) Then, from (8), C is a MRF with *local characteristics*

$$p_i(k|c_{S-i}) = \exp[\alpha(k) + \beta H_i(k, c_{\partial i})] / \sum_{c_i=1}^K \exp[\alpha(c_i) + \beta H_i(c_i, c_{\partial i})], \quad (9)$$

where $H_i(k, c_{\partial i})$ is the number of neighbours of i whose class labels agree with k . A high value of β gives a premium for homogeneous neighbourhoods, while β close to zero means nearly independent class labels. Negative β values are also possible for some phenomena, but probably not for remote sensing applications.

Now turn to applications of MRF models in image analysis. It is conceivable that at least the local behaviour of class labels over a scene is describable by an appropriate MRF (perhaps with a more complex clique-and-potentials structure than the one above). In the simplest case X_i 's are conditionally independent, and

$$\begin{aligned} p(c|x) &= \text{const.}(x) p(c) f(x|c) \\ &= \text{const.} \exp \left[\sum_i \{ \alpha(c_i) + \log f_{c_i}(x_i) \} + \beta \sum_{i < j} V_{i,j}(c_i, c_j) \right]. \end{aligned}$$

Accordingly C given data is again a MRF, only with updated $\alpha(c_i)$'s. Geman and Geman (1984) provide an ingenious “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 (6)). 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 computational 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 $\bar{c}$ for the scene, and then update $\bar{c}$ to a perhaps better $\hat{c}$ by finding $\hat{c}_i = k$ to maximise

$$p_i(C_i = k|x, \bar{c}_{S-i}) = \text{const.} \exp[\alpha(k) + \beta H_i(k, \bar{c}_{\partial i})] f_k(x_i). \quad (10)$$

(Note that this is like ordinary discriminant analysis but with neighbour-influenced “prior” probabilities $\pi(k) = \text{const.} \exp[\alpha(k) + \beta H_i(k, \bar{c}_{\partial i})]$.) In this way the full scene is swept over, in some order, and we have a new, updated estimate $\hat{c}$. The process is iterated until convergence; usually 6-10 times suffice. The starting point is ordinarily that corresponding to $\beta = 0$, i.e. the noncontextual classifier of 2A, 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 X -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:

$$f(x|c) = \text{const.} \exp -\frac{1}{2} \left[\sum_i (x_i - \mu_{c_i})' \Sigma^{-1} (x_i - \mu_{c_i}) - \sum_{i < j} \gamma_{ij} (x_i - \mu_{c_i})' \Sigma^{-1} (x_j - \mu_{c_j}) \right],$$

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

$$f(x_i|c, x_{S-i}) = N\{\mu_{c_i} + \gamma(\bar{x}_{\partial i} - \bar{\mu}_{\partial i}), \Sigma\}(x_i), \quad (11)$$

writing $\bar{x}_{\partial i}$ for the average of the four immediate x_j neighbours to x_i and similarly $\bar{\mu}_{\partial i}$ for the average of the four accompanying μ_j 's. It is easily seen that C given x again is a MRF. The Geman and Geman method is capable of coming close to the simultaneous optimisation of $p(c|x)$, i.e. of maximising

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

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 $\bar{c}$ of the rest of the scene,

$$p_i(C_i = k|x, \bar{c}_{S-i}) = \text{const.} \exp[\alpha(k) + \beta H_i(k, \bar{c}_{\partial i})] f(x_i|k, \bar{c}_{S-i}, x_{S-i}). \quad (12)$$

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

2D. Other approaches to classification

Let us briefly touch upon some other spatial classification techniques.

Underlying the methods of 2B and 2C are the loss functions (1) and (6). A simple intermediate loss function that is less crude than (6) but nevertheless has a “contextual element” is L = number of misclassified 2×2 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. To illustrate the gain over the noncontextual rule, one may evaluate the probability of successfully recognising a full 2×2 block, i.e. classifying each of the four pixels correctly. This probability is $\Phi(\delta/2)^4$ for the noncontextual rule in the two-class, Gaussian case with equal covariance matrices, where δ is the Mahalanobis distance. The figure below illustrates the kind of improvement one can achieve by a spatial model. Details are provided in Hjort (1987).

***** Nils-figure 2 here!!! *****

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. Although some of these originally were proposed without statistical modelling intentions, they are now perhaps best understood in such terms: they are generally related to the schemes described in 2C and to MRF models. A general view is taken in Kay and Titterton (1986), where connections to statistical analysis of incomplete data, and therefore quickly to EM-type algorithms, also are pointed out. The theory is still lacking in completeness, e.g. with respect to convergence proofs. A study of several relaxation algorithms is provided by Fiskum (1986).

An important problem in many applications is that of *segmenting* the image into homogeneous connected regions. A more probable-looking image than the original is produced, and a “region classifier” can be put to work afterwards. Spatial segmentation is important in situations where there is no set of a priori well defined classes to allocate pixels to; perhaps such classes can be defined only after an initial explorative analysis involving segmentation. — A variant of the block-method described above may be used as a step in segmentation. One can use the posterior probability of a block being homogeneous, i.e. all four pixels are from the same class, as a measure of its homogeneity, and put up a decision rule for determining which blocks are “border blocks” and which are not.

Of course there are many other approaches to spatial segmentation. Cross and Mason (1985) provide a split-and-merge method that employ local measures of texture as well as the spectral vectors. Also, many standard clustering techniques can be made “spatial” by including pixel location in some way in the distance measure used. The model-based approach seems to be most powerful, though. Several of the classification rules described in 2B can be used as spatial clustering schemes too, with appropriate starting points from perhaps tiny and artificial training sets, and possibly after an initial bout of “automatic, unsupervised training”, a method briefly discussed in 4B. Finally MRF models like in 2C can be used for segmentation, see Derin, Elliott, Cristi, and D. Geman (1984) and Elliott and Derin (1984).

A useful approach for improving on the noncontextual rule is to augment one’s feature vectors with further pertinent components (and stick to the noncontextual algorithm using the extended vectors). Adding an average of the neighbour vectors will usually give lower pixelwise error rate, and has a basis in a reasonable statistical model (for the X ’s), as Switzer (1980) has shown. One might also consider using texture-type information measures computed from neighbourhoods. ContextVision’s GOP 300 IMAGE PROCESSOR has available as standard variables pairs of filters giving “local orientation” and “local frequency”, and the use of these for classification is explored to some degree in Conradsen and Nielsen (1987). Finally we mention the possibility of using geographical information as auxiliary components. This requires ties from satellite image machinery to suitable data-bases, making up a Geographical Information System, and has recently become a reality in several countries.

There are various ways of pre-processing the image data, in order to achieve better

classification with the transformed data, or simply and often as importantly to enhance the visual quality and clarity. Using principal component analysis (or the Karhunen-Loève transform) is quite popular, the idea being that if the “signal” varies much more than the “noise”, then a grey-level image showing the first principal component (the linear combination with maximal variation) will display mostly signal. There are better ways that take the spatial nature of the problem into account. Switzer and Green (1984) propose a set of min/max autocorrelation factors that aim at isolating noise from signal, and the principal component analysis emerges as a special case. Green, Berman, Switzer, and Craig (1987) generalise further to a maximum noise fraction transformation that has as output new components ordered by image quality. — These methods attack the whole multivariate image without regard to special issues like available training sets from known classes etc., i.e. work directly with the marginal mixture distribution of the X -vectors.

2E. Discussion

We have discussed several approaches to the problem of classifying pixels in predefined classes, in particular the “local” in 2B and the “global” in 2C.

The neighbourhood model approach needs only local models, not necessarily derived from global ones. On the other hand the global approach is currently limited to the use of MRFs, which are given in terms of their local characteristics, and it is primarily this local structure that is involved in Besag’s and Geman and Geman’s algorithms. This is probably just as well, since the large-scale behaviour of MRFs can be rather peculiar and is not sufficiently well understood. Realisations may for example consist of a single class over the whole scene, exhibiting long range dependence, see for instance Pickard (1987).

There are simple variations on the basic families of MRF models that can alleviate some of the problems connected to strange large-scale behaviour. If it is expected that classes $1, \dots, K$ exist on the scene in proportions about $\pi_0(1), \dots, \pi_0(K)$, say, then this fact should be modelled into the prior distribution. P. Green (1986) proposed modifying an original $p(c)$ distribution using

$$p_h(c) = \text{const. } p(c) \exp\{-hD(c)\}$$

for a perhaps small number h , where $D(c) = \sum\{\pi(k) - \pi_0(k)\}^2$ measures the distance from the proportions $\pi(k) = \sum I\{c_i = k\}/N$ on the scene c to the target values. The effect of this on Besag’s algorithm, see e.g. (10) and (12), is feasible, in spite of the fact that C and C given x not longer are MRFs. The end product is a scene more in accordance with the prior guesses $\pi_0(1), \dots, \pi_0(K)$ than else would have been obtained. — Note that this scheme provides a way of simulating MRFs conditional on some events having to take place, by constructing an appropriate $D(c)$ -distance and making h large; see Cross and Jain (1983) for techniques and applications.

In principle a global specification $p(c)$ of the distribution for C should produce a local model for a neighbourhood; however, it is usually notoriously impossible to derive marginal distributions for sets of variables for MRFs. Pickard’s (1977) product model, that led to the neat local rules (4) and (5), is an exception.

The local and global viewpoints came from the very different specifications of the loss functions; cf. (1) and (6). (6) might be rejected on the grounds of it being too

un-compromising in its all-or-nothing stand; however, the resulting global restorations are generally quite acceptable, and, perhaps, not that different from the locally restored ones. Derin, Elliott, Cristi, and D. Geman (1984), on the basis of extensive simulations in the one-dimensional case (where MRFs are nothing but Markov chains), noted that solutions based on the pixelwise criterion (1) performed “just as well, if not better”, than the simultaneous maximum a posteriori solution. In the same vein, Preston (1971) noted in an early paper (on one-dimensional problems again) that solutions under the global criterion (6) performed well under loss function (1) too.

Of course other loss functions can be put to use. We mentioned one intermediate choice for blocks of pixels at a time in 2D. A simpler but important variation on (1) is one that weights different classification errors differently; in some applications not detecting a rare class is worse than misclassifying every-day pixels. The mathematical apparatus is easily changed to accomodate this; instead of maximising posterior probability one has to minimise conditional expected loss.

In many image analysis problems it appears safe to assume white noise, i.e. the X -vectors are independent from pixel to pixel, given the class labels. But in remote sensing applications with small surface element area this is not the case. Some methods that take such X -dependence into account have been noted here, cf. (5) and (12), but all in all the potential of the described approaches is somewhat unexplored when it comes to effective dependence models for X -vectors given the class labels. We have some preliminary experience with different model specifications for the h -term in formula (3), and also with appropriate generalisations of the conditional auto-regressive scheme (11). A particular model that is both reasonably realistic and leads to computationally feasible algorithms for segmentation and classification is the following “*random level*” model: assume $X = S + \varepsilon = \text{signal} + \text{noise}$, where S is constant for each connected region of class k , but is allowed to vary from k -region to k -region. Details pertaining to the analysis of this model will be given elsewhere.

A fact to be noted in these situations with autocorrelated X -data is that the contextual rules offer *less improvement* on the ordinary noncontextual rule, in terms of error rate, than they do in the case of white noise. This can be proved theoretically for several of the methods, and is apparent in the simulation study of Hjort, Mohn, and Storvik (1986): while the noncontextual error rate is unaffected by dependence structure, of course, the white-noise derived contextual rules have error rates that increase fast with correlation, while error rates for the contextual rules that explicitly take X -dependence into account also increase, but not so fast. The misclassifications tend to occur in patches near the boundary, and in some cases these patches stretch unreasonably far into homogenous land.

It may even happen that the naïve, noncontextual rule is best! This phenomenon is exhibited by Devore (1973) for a simple, discrete one-dimensional model. For a continuous example, assume (V_i, X_i) are Gaussian variables where X is observed and V is not. If V and X are processes with the same amount of correlation from site to site, as formalised by $\text{COV}\{(V_i, X_i), (V_j, X_j)\} = r(|j - i|) \Sigma$, $r(\cdot)$ scalar and $\Sigma = \text{VAR}\{(V_i, X_i)\}$, then the optimal predictor of V_i (and any function of V_i) is the naïve, noncontextual one. (We owe this last remark to Jon Helgeland.)

It transpires that one still needs more experimentation (and more theory) with mod-

els from several approaches to cover the fast-growing needs of disciplines that are now beginning to use image analysis as a tool.

3. Predicting a continuous variable

In several applications one is more interested in estimating a “ground parameter” than in classifying pixels into ground classes. Imagine that z_i is such a variable of interest, associated with ground surface element no. i . In a water quality application of MSS- and LANDSAT-satellite data we are working on there are in fact several z_i ’s of interest: the amount of plankton in sea element no. i , along with turbidity, water transparency, and other measures of water quality. In a SPOT-application to forestry the total “tree mass volume” for each 20m×20m element on the ground is of interest. — If one is lucky these z_i ’s are correlated with the remotely sensed X_i -vectors, and predictions based on these can be made. For example, several studies relating spectral satellite data to “ground truth” water quality variables indicate that such variables sometimes are approximately linearly related to certain chromaticity indices, these being (non-linear) functions of the spectral signatures (specifically, $ci(1) = x4/(x4 + x5 + x6)$ and $ci(2) = x5/(x4 + x5 + x6)$, where $x1, \dots, x6$ are data from the six channels of the MSS; see Lindell et al. (1985)). Other examples of this sort abound in the remote sensing literature, and make clear that many important surveillance tasks now can be carried out using remote sensing.

Let us for convenience still use X_i to denote the vector of pertinent observations at pixel no. i , for example spectral data (possibly transformed, as above) supplemented with other other available covariates thought to be useful for the prediction of z_i . In the water quality application, the components of X_i are the two chromaticity indices supplemented with topographic information (distance from land, for example). 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

$$\begin{pmatrix} Z_i \\ X_i \end{pmatrix} | (C_i = k) \sim N\left\{ \begin{pmatrix} \nu_k \\ \mu_k \end{pmatrix}, \begin{pmatrix} \tau_k^2 & \omega_k' \\ \omega_k & \Sigma \end{pmatrix} \right\}. \quad (13)$$

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

$$\widehat{z}_i(k) = E\{Z_i | x_i, C_i = k\} = \nu_k + \omega_k' \Sigma^{-1} (x_i - \mu_k), \quad (14)$$

and the conditional variance

$$\sigma(k)^2 = \text{Var}\{Z_i | x_i, C_i = k\} = \tau_k^2 - \omega_k' \Sigma^{-1} \omega_k \quad (15)$$

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 $P_i\{k|x_i\}$ to form a (non-contextual) predictor. Let us now examine some contextual alternatives.

3A. Predictors constructed from neighbourhood models

Assume that Z_i ’s given X and C are independent, with

$$Z_i | (x, c) \sim f(z_i | x_i, c_i) = N\{\widehat{z}_i(c_i), \sigma(c_i)^2\} (z_i). \quad (16)$$

A natural measure of the loss incurred by using $\hat{z}_i$ for z_i and $\hat{c}_i$ for c_i , $i = 1, \dots, N$, is

$$L[(c, z), (\hat{c}, \hat{z})] = \frac{1}{N} \sum_{i=1}^N I\{\hat{c}_i \neq c_i\} + \frac{1}{N} \sum_{i=1}^N (\hat{z}_i - z_i)^2$$

(or with the two terms differently weighted). Then the risk, or expected loss, is minimised by choosing $\hat{c}_i$ to maximise $P_i\{k|\text{data}(i)\}$, just as in Section 2B, and $\hat{z}_i = E\{Z_i|\text{data}(i)\}$. Reasonable approximations, and optimal solutions themselves within those that only utilise a subset $\text{data}(i)$ of the data to decide on c_i and z_i , are

$$\begin{aligned} \hat{c}_i &= \text{the one maximising } P_i\{k|\text{data}(i)\}, \text{ as before,} \\ \hat{z}_i &= E\{Z_i|\text{data}(i)\} = \sum_k P_i\{k|\text{data}(i)\} \hat{z}_i(k). \end{aligned} \tag{17}$$

One readily shows that Z_i given $\text{data}(i)$, i.e. the information from pixel i and its neighbours, is distributed as a mixture of the densities (16) with $P_i\{k|\text{data}(i)\}$ as weights. Accordingly,

$$\text{Var}\{Z_i|\text{data}(i)\} = \sum_{k=1}^K P_i\{k|\text{data}(i)\} \{\sigma(k)^2 + (\hat{z}_i(k) - \hat{z}_i)^2\},$$

which is of use when confidence intervals are called for. Observe that if a contextual classification is carried out using the “local modelling” method of (3), say corresponding to formulae (4) or (5), then the extra computational burden needed to compute predictors and standard deviations is very mild.

A somewhat different loss function that could be used is

$$L[(c, z), (\hat{c}, \hat{z})] = \frac{1}{N} \sum_{i=1}^N I\{\hat{c}_i \neq c_i \text{ or } |\hat{z}_i - z_i| > \varepsilon\},$$

for a small number ε . The optimal solution then is to use, for each pixel, the *mode* in the simultaneous conditional distribution of (C_i, Z_i) given the data. The neighbourhood approximation version of the optimal rule uses the fact that

$$(C_i, Z_i)|\text{data}(i) \sim P_i\{k|\text{data}(i)\} f(z_i|x_i, k),$$

and is different from (17) above: find first k to maximise $P_i\{k|\text{data}(i)\}/\sigma(k)$, then use this k for $\hat{c}_i$, and use $\hat{z}_i(k)$ for $\hat{z}_i$. We hasten to add, however, that the first loss function, which led to (17), is judged more reasonable in practice. The second loss function is not unreasonable either, but was included partly to show that going for the mode of the posterior distribution sometimes has unexpected consequences: in this case the mere fact that z -prediction also was of interest changed also the procedure for estimating c .

3B. Markov random field methods

Assume that C is a MRF with distribution (9), that X given c is a MRF with local characteristics (11), and that (16) holds regarding the distribution of Z given c and x . Then one can show that (X, Z) has a simultaneous MRF distribution given c , and that

(C, X, Z) is a simultaneous MRF. The most important fact is that the unobserved processes (C, Z) given the data $X = x$, the Bayesian posterior distribution, is a MRF too, with local characteristics

$$(C_i, Z_i)|(x, c_{S-i}, z_{S-i}) \sim \text{const.} \exp[\alpha(c_i) + \beta H_i(c_i, c_{\partial i})] f(x_i|c, x_{S-i}) f(z_i|x_i, c_i), \quad (18)$$

cf. (11) and (16). This generalises the result (12).

Besag's method of taking iterative conditional modes can be generalised to the present state of affairs, with both c and z to be estimated. Let $\bar{c}$ and $\bar{z}$ be initial estimates, say based on non-contextual analysis. Then update $\bar{c}_i$ and $\bar{z}_i$ to $\hat{c}_i$ and $\hat{z}_i$ by maximising (18) jointly, holding $\bar{c}$ and $\bar{z}$ fixed for the rest of the pixels. This amounts to finding k to maximise

$$\exp[\alpha(k) + \beta H_i(k, \bar{c}_{\partial i})] f(x_i|k, c_{\partial i}, x_{S-i})/\sigma(k),$$

see (15), use this k for $\hat{c}_i$, and finally use $\hat{z}_i(k)$ of (14) for $\hat{z}_i$. In particular the method for constructing $\hat{c}$ is not the same as the ordinary Besag method.

Again the (this time implicitly) chosen loss function is the reason for this somewhat unexpected result. The conditional mode method above is really aiming at optimal performance w.r.t.

$$L[(c, z), (\hat{c}, \hat{z})] = \begin{cases} 0 & \text{if every pixel is correctly recognised} \\ & \text{and } |\hat{z}_i - z_i| < \varepsilon \text{ for every } i; \\ 1 & \text{otherwise.} \end{cases}$$

In applications of this sort it is more important to come close to z_i most of the time than trying to hit exactly (which is impossible), and a more sensible loss function is $I\{\hat{c} \neq c\}$ plus the average squared prediction error. Under this loss structure the Bayes solution looks more natural: first use the usual Besag method (in this case, as explained around (12)) to arrive at the scene estimate $\hat{c}$. Then compute

$$\begin{aligned} \hat{z}_i &= E\{Z_i|x, \hat{c}_{S-i}, z_{S-i}\} \\ &= \frac{\sum_{k=1}^K \exp[\alpha(k) + \beta H_i(k, \hat{c}_{\partial i})] f(x_i|k, \hat{c}_{S-i}, x_{S-i}) \hat{z}_i(k)}{\sum_{k=1}^K \exp[\alpha(k) + \beta H_i(k, \hat{c}_{\partial i})] f(x_i|k, \hat{c}_{S-i}, x_{S-i})}. \end{aligned}$$

Note that the autocorrelation parameter γ enters via $f(x_i|c, x_{S-i})$.

In this case $\hat{z}_i$ emerged as an explicit function of the finally classified scene $\hat{c}$. This is because we assumed (16) to hold. In models where Z_i 's must be taken inter-dependent, given scene and X -data, a simultaneous, iterational updating of c and z may be called for.

4. Estimation problems

What we perhaps somewhat silently have subscribed to in the previous sections is nothing but the Bayesian paradigm: a collection c of parameters are unknown to us, but they have a prior distribution $p(c)$; we observe data X with likelihood $f(x|c)$ which leads to the posterior distribution $p(c|x)$; and inference about c , in particular estimation of the c_i 's, is determined from the posterior in tandem with the loss function. It is the fact that

the unknown c_i 's (along with the unknown z_i 's, in Section 3) are so likely to exhibit some natural structure that leads us to the Bayesian viewpoint, as opposed to for example the framework of compound decision problems. — But there are complications: our models have unknown parameters, and they must be estimated in some way. This brings us back to the more frequentist-like framework of empirical Bayes estimators.

The paragraph above used $p(c)$ and $f(x|c)$, but the reasoning applies equally well to the “local” framework, in which parallels are found in $\pi(k)$ and $g(\cdot|k)$ (the prior) and $h(\text{data}(i)|\cdot)$ (the likelihood); see formula (3). — Let us in general terms write $p(c) = p(c; \phi)$ and $f(x|c) = f(x|c; \theta)$ to signalise parametric structure. Thus ϕ could consist of $\alpha(1), \dots, \alpha(K)$, β , while θ could be the collection of class means and covariance matrices, and the correlation parameter γ , cf. Section 3B.

4A. Traditional supervised training

In remote sensing applications it is customary to have an expert defining (possibly small) training sets for the classes in question. The training set for class k is a set of X -vectors from pixels that are guaranteed to come from the class, and typically stem from a small number of homogeneous regions. This makes it possible to estimate class parameters in $f_k(x)$, say mean and covariance matrix. It is also possible to estimate auto-correlation parameters using such data. The parameter γ involved in (11), for example, can be estimated either by full maximum likelihood or by maximum pseudo-likelihood, i.e. the product of terms (11) over all sites within class k is maximised. (In this case it amounts to a sort of “conditional least squares”.) — Parameters involved in (13), necessary in order to estimate the co-variable z_i , can also be estimated based on such training areas, provided z -data also are collected in situ. This is usually a costly affair, but is often done when new types of remote sensing applications are explored, for example in order to establish functional relationships with the ground parameters of interest.

What is considerably more difficult than defining “homogeneous training areas” as above is to obtain ground truth for a representative region of land where the classes vary; again it can sometimes be necessary for experts to explore the region in situ. In such cases also parameters ϕ in the distribution of classes can be obtained in traditional ways. Direct maximum likelihood estimation is typically very difficult, but approximations are still computationally possible, through computer sampling to estimate the normalising constant for different parameter values, or to estimate expectation and covariance matrix of the sufficient statistics under a first-stage estimate, see Künsch's suggestion (1986). In many cases maximum pseudo-likelihood will be a more suitable procedure, in spite of lower efficiency in the idealised but unrealistic case that the model is exactly correct. The two methods really take aim at two different parameter values; see comments in Hjort (1987b), who also studies behaviour and efficiency of the pseudo-likelihood procedure in the simplest discrete case, that of (one-dimensional) Markov chains.

4B. Automatic, unsupervised training

Assume that the supervised training phase alluded to above gave mean vectors and covariance matrices for the K classes, assuming a Gaussian model for each class. These estimates may not be quite satisfactory, because of possibly too small training sets, and

because of positive correlation within each homogeneous training set, i.e. the variability is easily underestimated, with consequences for later classification. There is a pleasant way in which to exploit all the unclassified vectors, however. Pick a set of such vectors, say $X_1, \dots, X_M$, lying sufficiently far from each other to make independence between them a safe assumption. These, provided they pass an outlier test, come from the mixture density $f(x_i) = \sum_k \pi(k) f(x_i|k)$, and carry information about its parameters. The full likelihood for training data and the set of unclassified data can be written up, and maximisation w.r.t. means, covariance matrices, and prior probabilities $\pi(k)$ can be carried out with the EM algorithm, with starting values from the initial training sets. Details about this and related procedures are in Hjort (1985a, 1986). — The set of distant pixels chosen above can next be moved to a new set, etc., and eventually an average should be made to define these new, “automatically updated” parameter estimates.

Hjort (1985a) takes this approach further, by sampling clumps of neighbourhood data, say $\text{data}(1), \dots, \text{data}(M)$ being X -vectors from each of M pixel crosses. The likelihood for one such $\text{data}(i)$ collection of five vectors is

$$f(\text{data}(i)) = \sum_{k,a,b,c,d} \pi(k) g(a,b,c,d|k) h(\text{data}(i)|k,a,b,c,d),$$

see (3) once more. For given models $g(\cdot)$ and $h(\cdot)$ the parameters describing them can be estimated with maximum likelihood. — In this way it becomes possible to estimate for example the neighbour transition probabilities $p(a|k)$, involved in (4) and its generalisations, without further ground truth! It is also possible to estimate ϕ -parameters of $p(c; \phi)$ using this technique, for example Besag’s β .

4C. Area estimates

The true proportion of pixels from class k on the scene is $\pi(k) = \sum_{i=1}^N I\{c_i = k\}/N$ (assuming outliers removed from the N). The Bayes estimators under quadratic loss are

$$\pi(k) = E\{\pi(k)|\text{data}\} = \frac{1}{N} \sum_{i=1}^N P_i\{k|\text{data}\} \approx \frac{1}{N} \sum_{i=1}^N P_i\{k|\text{data}(i)\},$$

and these are much better than just counting occurrences in a classified image. — The estimator above is easiest to use under the “local model” approach of 2B, but they can also be computed under MRF type priors, through simulations from the posterior distribution.

4D. Simultaneous estimation

The idea of simultaneously estimating θ and ϕ , and use these estimates to produce the final picture $\hat{c}$, is certainly attractive, since it would be completely automatic (perhaps apart from initial guesses for class parameters). The task is difficult, however. Besag (1986) proposes the following: (i) use initial guess θ_o and a noncontextual rule (or one corresponding to a ϕ_0) to get $\hat{c}_0$; (ii) put your preliminary trust in $\hat{c}_0$ and obtain $\hat{\theta}_1$ by (e.g.) maximising $f(x|\hat{c}_0; \theta)$; (iii) still trusting $\hat{c}_0$, use maximum-PL to obtain $\hat{\phi}_1$; (iv) obtain a new scene estimate $\hat{c}_1$ by (e.g.) a single-cycle conditional mode; and (v) iterate.

There are difficulties with this scheme, stemming from the trust put in $\hat{c}_0$, and *it is not consistent*. This is easiest to see in the simple case of independency between pixels,

corresponding to $\beta = \phi = 0$, bringing us to the following problem: how to estimate class parameters based on independent, unclassified feature vectors. Besag's scheme reduces to what has been termed the "decision directed method" in pattern recognition literature. It is *biased*; a better method employs maximum likelihood on the mixture distribution, ending up in EM-iteration equations of the type $\mu_k = \sum_i P(k|X_i) X_i / \sum_i P(k|X_i)$ involving posterior probabilities (see e.g. Hjort 1986, Ch. 7). (i) in Besag's method corresponds to truncating these to 0 and 1. — We therefore recommend that class parameters θ get estimated by choosing grids of distant vectors and the mixture-density based maximum likelihood method, this (heavy) step replacing (i) above. Similarly step (ii) can be polished by taking uncertainty of $\hat{c}_0$ into account.

5. Further topics

We conclude our presentation of remote sensing statistics with mentioning a couple of other issues of great interest.

5A. Mixed pixels

That the assumption of a unique, well-defined class for each pixel may be too crude has been mentioned already. The trick of using the "doubt" option (see (1) and (2)) often ameliorates the problem: in many cases the mixed pixels occur along boundaries, and the pixels classified as doubts often lie exactly there, in our experience. Furthermore, the estimated posterior probabilities that are involved in the classification rules provide of course estimates of the proportions of a pixel that stem from the various classes. Another and more serious method of systematically handling such un-pure pixels is to model the proportions themselves as a stochastic process. This is what Kent and Mardia (1986) do in their "fuzzy" classification. They employ a multivariate logistic autonormal MRF process to model the fuzzy membership process, and use concepts and techniques resembling those of Sections 2C and 3C here. Finally we mention Switzer's work (1983) on this topic. He proposes estimates for the class-proportions of a pixel, assuming locally-linear boundaries.

5B. Images repeated in time

Assume the same scene of land is covered by satellite at time points t and t' , producing collections of data x and x' . Assuming the geometrical correction problems to be solved properly we assume that the second image really is "on top of" the first one, so that the underlying classes, say c_i and c'_i for pixel no. i , refer to the same ground area element. — Interest often focusses on whether changes have occurred on the ground, and one might be specifically interested in detecting all the pixels that have changed. Assume that the class distributions are the same for the two points of time (this could mean that the grand mean for the scene has been subtracted out in each image). A simple simultaneous MRF model for (C, C') is

$$q(c, c') = \text{const.} \exp \left[\sum_i \gamma(c_i) I\{c_i = c'_i\} \right] p(c) p(c'),$$

where $p(\cdot)$ is the MRF distribution for one scene alone. A high value of $\gamma(k)$ means that it is very likely that pixel i does not change, whereas $\gamma(k)$ near zero means that if $c_i = k$,

then c'_i behaves nearly independently of c_i . — Now a procedure for detection of all changed pixels can be put up in a way resembling the Besag method of Section 2C. Similarly the neighbourhood modelling approach can be taken; procedures for detecting changed pixels are in Mohn (1987).

The model $q(c, c')$ above suggests that changes occur randomly over the scene. If prior knowledge indicates that one pixel's changing encourages its neighbour's change too, then another model should and could be put up. We make this point partly to illustrate that these challenging problems might be very application-dependent, and we cannot very likely hope to construct methods that perform well in any situation.

5C. Detecting structure

There are various image analysis type techniques for enhancing one-dimensional structure in the two-dimensional picture. The problem of finding roads automatically from satellite data, for example, is of this type, and can also be attacked with statistical methods, employing carefully modelled MRFs, for example. It is important to realise that these problems typically require separate and application-dependent solutions, i.e. one cannot rely solely of a good contextual classification rule to automatically find structures involving several pixels.