

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

CONTEXTUAL CLASSIFICATION OF REMOTELY SENSED DATA,
BASED ON AN AUTO-CORRELATION MODEL

by

Nils Lid Hjort, Erik Mohn, and Geir Storvik
Norwegian Computing Center

TECHNICAL REPORT NO. 13
NOVEMBER 1985

PREPARED UNDER THE AUSPICES
OF
NATIONAL SCIENCE FOUNDATION
GRANT MCS 81-09584

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA



**CONTEXTUAL CLASSIFICATION OF REMOTELY SENSED DATA,
BASED ON AN AUTO-CORRELATION MODEL**

by

**Nils Lid Hjort, Erik Mohn, and Geir Storvik
Norwegian Computing Center**

**TECHNICAL REPORT NO. 13
NOVEMBER 1985**

**PREPARED UNDER THE AUSPICES
OF
NATIONAL SCIENCE FOUNDATION
GRANT MCS 81-09584**

**DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA**

CONTEXTUAL CLASSIFICATION OF
REMOTELY SENSED DATA, BASED ON
AN AUTO-CORRELATION MODEL

by

Nils Lid Hjort, Erik Mohn, Geir Storvik
Norwegian Computing Center
Oslo, Norway

ABSTRACT

It has been usual when designing classification procedures for remote sensing data to assume independence of the feature vectors. This assumption may be reasonable when the pixels are large, say 100 x 100 m, but less realistic for high resolution satellites with pixel sizes of 20 x 20 m, or less.

This paper considers a simple model which allows for spatial correlation. The idea is to express the feature vectors from neighbouring pixels as a sum of two independent processes, one having independent vectors with class dependent distributions, the other being a contaminating, auto-correlated noise process. This model is combined with different models for the joint distribution of the classes of the neighbours, giving contextual classification rules incorporating the spatial relationship of the feature vectors.

A method for estimating the parameters in the autocorrelation model is also presented.

1. INTRODUCTION

A scene consists of N pixels. Observed in pixel i is a d -dimensional vector X_i , the components of which are spectral measurements in different channels. X_i may also contain components bearing for example topographic information. K classes, having prior probabilities $\pi(k)$, $k = 1, \dots, K$, are defined on the scene. The X 's from class k are assumed to be distributed according to a density $f_k(x)$, in the present paper further assumed to be the Gaussian one.

Let C_i be the true class of pixel i . The aim of a statistical classification procedure is to estimate the C_i 's on the basis of the observed X_i 's. Most procedures require some preliminary knowledge about the class distributions, and sometimes also some further statistical parameters. This knowledge, for example in the form of parameter estimates, is usually extracted from a training set for each class and possibly further samples of 'ground truth'.

It is supposed in the present paper that all outliers, i.e. pixels not belonging to any of the K predefined classes, have been taken care of separately, such that each $C_i \in \{1, \dots, K\}$. (A procedure managing this is described in Hjort (1985a).) Let C_i^* denote the class assignment, according to a classification procedure, of pixel i . The possibility of being in doubt and stating nothing about C_i is reserved, having in mind, for example, pixels made up of a mixture of points from more than one class. Thus each $C_i^* \in \{1, \dots, K, D\}$, where D means 'being in doubt'.

Assume that the loss incurred when pixel i is assigned to C_i^* is of the type

$$L(C_i, C_i^*) = \begin{cases} 0 & \text{if } C_i^* = C_i \quad (\text{correct decision}) \\ 1 & \text{if } C_i^* \neq C_i \text{ and } C_i^* \in \{1, \dots, K\} \quad (\text{wrong decision}) \\ c & \text{if } C_i^* = D \quad (\text{being in doubt}) \end{cases}$$

where $c \in [0, 1 - (1/K)]$ is a certain threshold. Number the pixels that are to be classified $1, \dots, M$ (some of the original N pixels are used in training sets, and some are already classified as outliers). As a total

measure of the consequences involved when pixels with true classes $(C_1, \dots, C_M)$ are classified as $(C_1^*, \dots, C_M^*)$, we use

$$L_{\text{total}} = (1/M) \sum_{i=1}^M L(C_i, C_i^*),$$

which is seen to equal the fraction of misclassified pixels plus c times the fraction of 'being in doubt' pixels. The expected loss may be expressed as

$$EL_{\text{total}} = e^*(c) + cd^*(c)$$

where e^* and d^* are the error rate and doubt rate, respectively.

One may now show that the optimal (Bayes) rule, i.e. the classification procedure minimising EL_{total} , is of the form

$$C_i^* = \begin{cases} D & \text{if } P_i(k|\Delta) < 1-c \text{ for all } k=1, \dots, K, \\ m & \text{if } P_i(m|\Delta) = \max_{k \leq K} P_i(k|\Delta) > 1-c. \end{cases} \quad (1)$$

The rule involves the posterior probabilities

$$P_i(k|\Delta) = \Pr\{C_i = k|\Delta\}, \quad (2)$$

where Δ (for 'data') means the full scene's observations. It is seen that the Bayes procedure is dependent upon the specification of the probability distribution of $(C_1, \dots, C_N, X_1, \dots, X_N)$. This being an enormous vector, the algorithm (1) will usually be impossible to use in practice.

The following procedure is proposed in Hjort & Mohn (1984) as an approximation to the Bayes rule (1):

$$C_i^* = \begin{cases} D & \text{if } P_i(k|\Delta_i) < 1-c \text{ for all } k=1, \dots, K, \\ m & \text{if } P_i(m|\Delta_i) = \max_{k \leq K} P_i(k|\Delta_i) > 1-c. \end{cases} \quad (3)$$

Here

$$P_i(k|\Delta_i) = \Pr\{C_i = k|\Delta_i\}, \quad (4)$$

and Δ_i is a suitably chosen set of observations, including at least $\{X_i\}$, to take into account when trying to classify pixel i . A natural choice, which will be employed in most of the present paper, is

$$\Delta_i = \{X_i, X_{iN}, X_{iE}, X_{iS}, X_{iW}\}, \quad (5)$$

that is, observations from the four immediate neighbours lying north, east, south, and west of a pixel are also used when this centre pixel is classified. In this sense, the rules (3) are contextual, and aim at better performance than the traditional rules of discriminant analysis that are currently used in most commercial systems, in that these use only $\Delta_i = \{X_i\}$.

It is not difficult to see that the posterior probabilities (4) may be expressed as

$$\begin{aligned} P_i(k|\Delta_i) &= P_i(k|X_i, X_{iN}, X_{iE}, X_{iS}, X_{iW}) \\ &= (1/f(\Delta_i)) \pi(k) \tau_{a,b,c,d} g(a,b,c,d|k) f(\Delta_i|k,a,b,c,d). \end{aligned} \quad (6)$$

Here (a,b,c,d) is one of the K^4 possible configurations of neighbour classes, and $g(a,b,c,d|k)$ is the probability of seeing this configuration, given that the centre pixel is from class k . Further $f(\Delta_i|k,a,b,c,d)$ is the conditional joint density of the five vectors, given that k,a,b,c,d are the classes in question. The denominator $f(\Delta_i)$ is the unconditional density of the five spectral vectors, making $\sum_{k=1}^K P_i(k|\Delta_i) = 1$. Assuming that the conditional distribution of X_i , given the classes in the neighbourhood, just depends on the class of pixel i , the class density $f_k(X_i)$ may be factorised out in $f(\Delta_i|k,a,b,c,d)$ to get

$$f(\Delta_i|k,a,b,c,d) = f_k(X_i) h(X_{iN}, \dots, X_{iW}|X_i, k, a, b, c, d). \quad (7)$$

A general formula for $P_i(k|\Delta_i)$ results:

$$P_i(k|\Delta_i) = \text{const. } \pi(k) f_k(X_i) R_k(\Delta_i), \quad (8)$$

where

$$R_k(\Delta_i) = \int_{a,b,c,d} g(a,b,c,d|k) h(X_{iN}, \dots, X_{iW} | X_i, k, a, b, c, d) \quad (9)$$

acts as a correction or adjustment factor, taking spatial dependence into account, to the classical non-contextual method, which has

$$P_i(k|X_i) = \text{const. } \pi(k) f_k(X_i).$$

The general formulae (6) and (8) were presented and further commented upon in Hjort & Mohn (1984). The approach taken is seen to give rise to a variety of contextual classification methods, depending on how g and f in (6) (or equivalently, g and h in (9)) are modelled. Several published proposals may be seen to be special cases. Most of these assume conditional independence between neighbouring spectral vectors, given the classes, that is,

$$f(\Delta_i | k, a, b, c, d) = f_k(X_i) f_a(X_{iN}) \dots f_d(X_{iW}), \quad (10)$$

or

$$h(X_{iN}, \dots, X_{iW} | X_i, k, a, b, c, d) = f_a(X_{iN}) \dots f_d(X_{iW}),$$

and concentrate instead on constructing good models for the distribution of classes, leading to specifications for g in (9). Two references here are Owen & Switzer (1982) and Haslett (1983).

It is argued in Section 2, however, that the conditional independence assumption (10) becomes inadequate when applied to small pixels, i.e. data from high resolution satellites. The present paper concentrates on a model for the spectral vectors that allows positive correlation between spectral measurements from neighbouring pixels, given the classes. This model is described in Section 2, and results in a specification of h in (9). When this h is combined with any choice of g , a contextual classification procedure arises.

The two mosaic models used by Haslett and Owen & Switzer (op.cit.) are singled out for special attention. In particular it is shown that a simple approximation to the rule combining Haslett's Markov random field and the correlation model of Section 2 produces a relatively simple and easily implementable algorithm, that incorporates both types of contextual information: transition probabilities for the classes and spectral spatial correlation.

The procedures emerging in Section 2 employ an auto-correlation parameter ρ and a proportion parameter θ . Section 3 offers methods to obtain estimates of these parameters.

In Section 4 some possible generalisations are briefly considered.

The proposed classification procedures are presently implemented on a VAX 11/780. They will be tried out and compared with other methods on artificial data as well as on a simulated SPOT-scene. Results will be reported in a forthcoming paper.

2. A MODEL FOR SPECTRAL OBSERVATIONS CONTAMINATED WITH AUTO-CORRELATED NOISE

One of the simplest statistical models for the behaviour of spectral vectors $(X_1, \dots, X_M)$ given the true classes $(C_1, \dots, C_M)$ is the following: the X_i 's are conditionally independent, and $X_i \sim N(\mu_{C_i}, \Sigma)$, $i=1, \dots, M$.

μ_k is envisaged as the mean of a large number of X_i 's from class k . A more realistic model is presumably one where the process governing the X_i 's is thought of as being the sum of two independent processes, say $(Y_1, \dots, Y_M)$ having the simple class-dependent structure above, and a contaminating, auto-correlated noise process $(\epsilon_1, \dots, \epsilon_M)$. We think of the Y_i 's as being the ones that directly bear information of a pixel's class, whereas the ϵ_i 's are imagined as being the result of measurement errors and possibly other sources of 'extra variation'.

To be specific, let $X_i = Y_i + \epsilon_i$, where $Y_1, \dots, Y_M$, $(\epsilon_1, \dots, \epsilon_M)$ are independent, $Y_i \sim N(\mu_{C_i}, (1-\theta)\Gamma)$, and where $(\epsilon_1, \dots, \epsilon_M)$ is multinormally distributed with zero means and

$$\text{COV}(\epsilon_i, \epsilon_j) = \rho^{|j-i|} \theta \Gamma. \quad (11)$$

In particular

$$\text{VAR} \epsilon_i = \theta \Gamma.$$

It follows that $X_i \sim N(\mu_{C_i}, \Gamma)$ given C_i , as in the simple model mentioned above. The important difference is that now the X_i 's are positively correlated:

$$\text{COV}(X_i, X_j) = \rho^{|j-i|} \theta \Gamma, \quad i \neq j \quad (12)$$

given the classes C_i, C_j . ($|j-i|$ means ordinary distance between (centres of) pixels in an integer-valued coordinate system.)

If $\theta = 0$, then the simple model mentioned first results (and likewise if $\theta > 0$ and $\rho = 0$). The other extreme possibility is $\theta = 1$, in which case X_i may be expressed as the class mean μ_{C_i} plus auto-correlated noise. This results in a model similar in spirit to the ones used by Kittler & Föglein (1984), and Mardia (1984). Kittler and Föglein's model is a Gaussian Markov field on the infinite rectangular lattice, considered previously by Besag (1974). The conditional distribution of X_i given all other X_j 's is the multinormal with an expectation equal to a linear function of the X_j 's from the neighbourhood of pixel i . In order to satisfy the stationarity inherent in the model a normalisation of the observed feature vectors has to be done. The starting point in Mardia (1984) is "local spatial continuity", which means that all pixels in the neighbourhood are assumed to have the same class. The X_i 's are multinormal with a covariance structure corresponding to $\theta = 1$ in our model. The situation we have in mind, however, is one with positive, but perhaps small values of ρ and θ .

The value of the autocorrelation parameter depends on the pixel size. The larger the pixels, the smaller the value. To get a quantitative impression of the relationship let us consider a SPOT scene with resolution 20 x 20 m and a LANDSAT scene with resolution 100 x 100 m. Denote the autocorrelation parameters for these two resolutions by ρ and $\tilde{\rho}$ respectively. Denote further a feature vector from an arbitrary pixel in a SPOT scene by X_i and in a LANDSAT scene by $\tilde{X}_i$. For simplicity we suppose that $\tilde{X}_i$ is the mean of the X_j 's from the 25 SPOT-pixels within LANDSAT pixel no i :

$$\tilde{X}_i = \frac{1}{25} \sum_{j \in \mathcal{P}_i} X_j,$$

where $\mathcal{P}_i$ is the index set for the SPOT pixels in the LANDSAT pixel. Then

$$\text{COV}(\tilde{X}_i, \tilde{X}_{i+1}) = \frac{1}{25^2} \sum_{j \in \mathcal{P}_i} \sum_{k \in \mathcal{P}_{i+1}} \text{COV}(X_j, X_k) = \frac{1}{25^2} \left(\sum_{j \in \mathcal{P}_i} \sum_{k \in \mathcal{P}_{i+1}} \rho^{|j-k|} \right) \theta$$

$$\text{VAR}(\tilde{X}_i) = \frac{1}{25^2} \text{VAR} \left(\sum_{j \in \mathcal{P}_i} X_j \right).$$

Requiring $\text{COV}(\tilde{X}_i, \tilde{X}_{i+1}) = \tilde{\rho} \theta \text{VAR}(\tilde{X}_i)$ we obtain after some algebra that

$$\tilde{\rho} = \frac{5A_0 + 8A_1 + 6A_2 + 4A_3 + 2A_4}{25 + \theta[5B_0 + 8B_1 + 6B_2 + 4B_3 + 2B_4]},$$

where

$$A_k = \sum_{j=1}^9 (5-|j-5|) \rho^{\sqrt{k^2+j^2}}, \quad k=0,1,2,3,4$$

$$B_0 = 2 \sum_{j=1}^4 (5-j) \rho^j$$

$$B_k = 5\rho^k + 2 \sum_{j=1}^4 (5-j) \rho^{\sqrt{k^2+j^2}}, \quad k=1,2,3,4.$$

With $\theta=1$ and $\rho=0.5$ we get $\tilde{\rho}=0.22$. The usual assumption of conditional independence of the spectral vectors may be acceptable for large pixels, but quite unrealistic for high resolution satellites. Estimates of ρ obtained in this paper from a SPOT scene of As, Norway, around 0.9 indicate that the independence assumption may be doubtful even for large pixels.

To construct a classification rule, we need to substitute in (6) an expression for $f(\Delta_i | k, a, b, c, d)$. This is a normal density in 5d dimensions:

$$\begin{bmatrix} X_{iN} \\ X_{iE} \\ X_{iS} \\ X_{iW} \\ X_i \end{bmatrix} \sim N_{5d} \left(\begin{bmatrix} \mu_a \\ \mu_b \\ \mu_c \\ \mu_d \\ \mu_k \end{bmatrix}, \begin{bmatrix} 1 & \beta 1 & \gamma 1 & \beta 1 & \alpha 1 \\ & 1 & \beta 1 & \gamma 1 & \alpha 1 \\ & & 1 & \beta 1 & \alpha 1 \\ & & & 1 & \alpha 1 \\ & & & & 1 \end{bmatrix} \right)$$

given that the classes are a, b, c, d, k . Here

$$\begin{aligned} \alpha &= \text{correlation between 1-neighbours} = \rho\theta, \\ \beta &= \text{correlation between } \sqrt{2}\text{-neighbours} = \rho^{\sqrt{2}}\theta, \\ \gamma &= \text{correlation between 2-neighbours} = \rho^2\theta. \end{aligned} \quad (13)$$

We need a suitable expression for the h-term in formulae (7) - (9), which is the joint conditional density of $(X_{iN}, X_{iE}, X_{iS}, X_{iW})$ given X_i , still given the classes.

By Anderson (1958, p.28), this density is normal with mean vector

$$\begin{bmatrix} \mu_a \\ \mu_b \\ \mu_c \\ \mu_d \end{bmatrix} + \begin{bmatrix} \alpha 1 \\ \alpha 1 \\ \alpha 1 \\ \alpha 1 \end{bmatrix} \Gamma^{-1} (X_i - \mu_k) = \begin{bmatrix} \mu_a + \alpha(X_i - \mu_k) \\ . \\ . \\ \mu_d + \alpha(X_i - \mu_k) \end{bmatrix} \quad (14)$$

and covariance matrix

$$\Gamma_* = \begin{bmatrix} (1-\alpha^2)\Gamma, & (\beta-\alpha^2)\Gamma, & (\gamma-\alpha^2)\Gamma, & (\beta-\alpha^2)\Gamma \\ & (1-\alpha^2)\Gamma, & (\beta-\alpha^2)\Gamma, & (\gamma-\alpha^2)\Gamma \\ & & (1-\alpha^2)\Gamma, & (\beta-\alpha^2)\Gamma \\ & & & (1-\alpha^2)\Gamma \end{bmatrix}. \quad (15)$$

The inverse of this matrix is needed. Define the 4 x 4 matrix

$$S = \begin{bmatrix} 1-\alpha^2, & \beta-\alpha^2, & \gamma-\alpha^2, & \beta-\alpha^2 \\ & 1-\alpha^2, & \beta-\alpha^2, & \gamma-\alpha^2 \\ & & 1-\alpha^2, & \beta-\alpha^2 \\ & & & 1-\alpha^2 \end{bmatrix}. \quad (16)$$

S has inverse of the form

$$S^{-1} = \begin{bmatrix} u, & v, & w, & v \\ & u, & v, & w \\ & & u, & v \\ & & & u \end{bmatrix}. \quad (17)$$

Because Γ_* can be written as the Kronecker product $\Gamma_* = S \otimes \Gamma$, it follows that $\Gamma_*^{-1} = S^{-1} \otimes \Gamma^{-1}$; thus

$$\Gamma_*^{-1} = \begin{bmatrix} u\Gamma^{-1}, & v\Gamma^{-1}, & w\Gamma^{-1}, & v\Gamma^{-1} \\ & u\Gamma^{-1}, & v\Gamma^{-1}, & w\Gamma^{-1} \\ & & u\Gamma^{-1}, & v\Gamma^{-1} \\ & & & u\Gamma^{-1} \end{bmatrix}. \quad (18)$$

Explicit expressions for u, v, w in terms of α, β, γ , and a fortiori in terms of ρ, θ , may be derived, but it is convenient for the moment to define the real numbers u, v, w simply by (16) and (17). The particular constant

$$t = u + 2v + w \quad (19)$$

will also be needed in what follows.

We may now proceed with the derivation of an expression for $h(X_{iN}, X_{iE}, X_{iS}, X_{iW} | X_i, k, a, b, c, d)$. Define

$$y_i(a, b, c, d) = y_i = \begin{bmatrix} X_{iN} - \mu_a \\ X_{iE} - \mu_b \\ X_{iS} - \mu_c \\ X_{iW} - \mu_d \end{bmatrix}, \quad z_i = \begin{bmatrix} X_i - \mu_k \\ X_i - \mu_k \\ X_i - \mu_k \\ X_i - \mu_k \end{bmatrix}; \quad (20)$$

Then

$$h(X_{iN}, \dots, X_{iW} | X_i, k, a, b, c, d) = (2\pi)^{-2d} |\Gamma_*|^{-1/2} e^{-1/2 Q_{i,k}(a, b, c, d)},$$

where

$$\begin{aligned} Q_{i,k}(a, b, c, d) &= (y_i - \alpha z_i)' \Gamma_*^{-1} (y_i - \alpha z_i) \\ &= y_i' \Gamma_*^{-1} y_i - 2\alpha y_i' \Gamma_*^{-1} z_i + \alpha^2 z_i' \Gamma_*^{-1} z_i \\ &= y_i' \Gamma_*^{-1} y_i - 2\alpha t(X_{iN} + \dots + X_{iW} - \mu_a - \dots - \mu_d)' \Gamma_*^{-1} (X_i - \mu_k) \\ &\quad + 4\alpha^2 t(X_i - \mu_k)' \Gamma_*^{-1} (X_i - \mu_k). \end{aligned}$$

Hence

$$\begin{aligned} \exp\{-\frac{1}{2} Q_{i,k}(a, b, c, d)\} &= e^{-2\alpha^2 t(X_i - \mu_k)' \Gamma_*^{-1} (X_i - \mu_k)} R_i(a, b, c, d) \cdot \\ &\quad e^{\alpha t(X_{iN} + \dots + X_{iW} - \mu_a - \dots - \mu_d)' \Gamma_*^{-1} (X_i - \mu_k)}, \end{aligned} \quad (21)$$

where

$$R_i(a, b, c, d) = \exp\{-\frac{1}{2} y_i(a, b, c, d)' \Gamma_*^{-1} y_i(a, b, c, d)\}. \quad (22)$$

Accordingly, by (8) and (9),

$$P_i(k|X_i, X_{iN}, X_{iE}, X_{iS}, X_{iW}) \\ = \text{const. } \pi(k) f_k(X_i) e^{-2\alpha^2 t(X_i - \mu_k)} \cdot \Gamma^{-1}(X_i - \mu_k).$$

$$\Gamma_{a,b,c,d} g(a,b,c,d|k) R_i(a,b,c,d) \exp\{\alpha t(X_{iN} + \dots + X_{iW} - \mu_a - \dots - \mu_d)\} \Gamma^{-1}(X_i - \mu_k) \} \quad (23)$$

where const. is independent of class k , and is determined by

$$\sum_{k=1}^K P_i(k|X_i, X_{iN}, \dots, X_{iW}) = 1.$$

Each particular choice of probability distributions $\{g(a,b,c,d|k)\}$ may now be used in (23) to produce a classification procedure of the form (3). The large number of terms involved in the general formula, K^4 , make it clear, however, that only choices of g that can lead to practical simplifications should be considered, unless the number of classes is very low.

Specification 1.

Many of the configuration probabilities $g(a,b,c,d|k)$ will be small, and the smallest ones could be set equal to zero, resulting in an approximation to (23). A simple geometric probability model, modestly generalising one by Owen & Switzer, (1982), is proposed by Hjort (1985c) and has $g(a,b,c,d|k)$ positive for only $8K-7$ configurations:

$$g(k,k,k,k|k) = p + (q+r)\pi(k),$$

$$g(k,k,m,m|k) = g(m,k,k,m|k) = g(m,m,k,k|k) = g(k,m,m,k|k) = \frac{1}{4} q\pi(m),$$

$$g(k,k,m,k|k) = g(k,k,k,m|k) = g(m,k,k,k|k) = g(k,m,k,k|k) = \frac{1}{4} r\pi(m),$$

$m \neq k$, for certain parameters p, q, r summing to unity. Then (23) simplifies (in the sense of number of terms, at least) to

$$\begin{aligned}
P_i(k|X_i, X_{iN}, X_{iS}, X_{iW}) \\
&= \text{const. } \pi(k) f_k(X_i) e^{-2\alpha^2 (X_i - \mu_k)' \Gamma^{-1} (X_i - \mu_k)} \\
&\quad [p \exp\{4\alpha t(\bar{X}_i - \mu_k)' \Gamma^{-1} (X_i - \mu_k)\} R_i(k, k, k, k) \\
&\quad + q \exp\{2\alpha t(\bar{X}_i - \mu_k)' \Gamma^{-1} (X_i - \mu_k)\} G_{i,k}(X_i, \bar{X}_i) \\
&\quad + r \exp\{3\alpha t(\bar{X}_i - \mu_k)' \Gamma^{-1} (X_i - \mu_k)\} H_{i,k}(X_i, \bar{X}_i)], \quad (24)
\end{aligned}$$

where

$$\begin{aligned}
G_{i,k}(X_i, \bar{X}_i) &= \sum_{m=1}^K \pi(m) \exp\{2\alpha t(\bar{X}_i - \mu_m)' \Gamma^{-1} (X_i - \mu_k)\} \\
&\quad \frac{1}{4} [R_i(k, k, m, m) + R_i(m, k, k, m) + R_i(m, m, k, k) + R_i(k, m, m, k)], \\
H_{i,k}(X_i, \bar{X}_i) &= \sum_{m=1}^K \pi(m) \exp\{\alpha t(\bar{X}_i - \mu_m)' \Gamma^{-1} (X_i - \mu_k)\} \\
&\quad \frac{1}{4} [R_i(k, k, m, k) + R_i(k, k, k, m) + R_i(m, k, k, k) + R_i(k, m, k, k)],
\end{aligned}$$

and

$$\bar{X}_i = \frac{1}{4} (X_{iN} + X_{iE} + X_{iS} + X_{iW}). \quad (25)$$

Specification 2.

The parameters p, q, r involved in the previous specification are defined in terms of probabilities for certain configurations of regions, see Hjort (1985c). It is seen that they depend on the pixel size, and in particular that for very small pixels (i.e. high resolution satellite data) compared to the usual region sizes, p becomes close to 1 and q, r become close to 0. Accordingly, we may propose this extremely special special case of specification 1, namely

$$\begin{aligned}
P_i(k|X_i, X_{iN}, X_{iE}, X_{iS}, X_{iW}) \\
= \text{const. } \pi(k) f_k(X_i) e^{-2t\alpha^2 (X_i - \mu_k)' \Gamma^{-1} (X_i - \mu_k)} \\
e^{4\alpha t (\bar{X}_i - \mu_k)' \Gamma^{-1} (X_i - \mu_k)} R_i(k, k, k, k), \quad (26)
\end{aligned}$$

as an approximation. This version is both easier to implement and should require less computer time than the one based on (24).

The assumption on g that leads to (26) is the rather extreme one that $g(k, k, k, k|k) = 1$, and corresponds to what Switzer (1980) and Mardia (1984) call local spatial continuity.

Specification 3.

Assume stationarity and isotropy, and let

$$\pi(m|k) = \Pr\{C_j = m | C_i = k\} \quad (27)$$

where j is one of the immediate neighbours to pixel i . Haslett (1983) and Hjort & Mohn (1984) consider a model having

$$g(a, b, c, d|k) = \pi(a|k)\pi(b|k)\pi(c|k)\pi(d|k), \quad (28)$$

$a, b, c, d = 1, \dots, K$. This model, also related to a particular Markov random field studied by Pickard (1980), was seen to lead to a considerable simplification of the formula for $P_i(k|X_i, X_{iN}, \dots, X_{iW})$ in the conditional independence case, see e.g. Hjort & Mohn (1984). A clever approximation for $R_i(a, b, c, d)$ in (22) is needed now, however, in order to simplify (23) under the model (28).

Define

$$R_{i,0}(a,b,c,d) = \exp\left\{-\frac{1}{2} y_i'(a,b,c,d) (\Gamma_{\star}^{-1})_0 y_i(a,b,c,d)\right\}, \quad (29)$$

where the matrix $(\Gamma_{\star}^{-1})_0$, while approximating $\Gamma_{\star}^{-1}$, is to be chosen so as to imply a multiplicative structure on $R_{i,0}(a,b,c,d)$. Let

$$(\Gamma_{\star}^{-1})_0 = \begin{bmatrix} u_0 \Gamma^{-1} & , & 0 & , & 0 & , & 0 \\ & u_0 \Gamma^{-1} & , & 0 & , & 0 \\ & & u_0 \Gamma^{-1} & , & 0 \\ & & & u_0 \Gamma^{-1} \end{bmatrix}.$$

It is reasonable to choose the constant u_0 so that the ellipsoids

$$y' \Gamma_{\star}^{-1} y \leq \text{const.}, \quad y' (\Gamma_{\star}^{-1})_0 y \leq \text{const.}$$

get the same volumes, having in mind that $R_{i,0}$ in (29) is to be near R_i in (22). In other words, the determinants

$$|\Gamma_{\star}^{-1}| = |S^{-1}|^d |\Gamma^{-1}|^4,$$

$$|(\Gamma_{\star}^{-1})_0| = u_0^{4d} |\Gamma^{-1}|^4$$

using (17) and (18), should agree. This leads to

$$u_0 = |S^{-1}|^{1/4} = |S|^{-1/4}.$$

Using this approximation we obtain the following simplification of (23). Substitute $R_{i,0}(a,b,c,d)$ there with

$$R_{i,0}(a,b,c,d) = \exp\left\{-\frac{1}{2} u_0 [(X_{iN}-\mu_a)' \Gamma^{-1} (X_{iN}-\mu_a) + \dots + (X_{iW}-\mu_d)' \Gamma^{-1} (X_{iW}-\mu_d)]\right\}$$

to get

$$P_i(k|X_i, X_{iN}, X_{iE}, X_{iS}, X_{iW}) = \text{const. } \pi(k) f_k(X_i) e^{-2\alpha^2 t(X_i - \mu_k)' \Gamma^{-1} (X_i - \mu_k)}$$

$$T_k(X_{iN}|X_i) T_k(X_{iE}|X_i) T_k(X_{iS}|X_i) T_k(X_{iW}|X_i), \quad (30)$$

where

$$T_k(x|x_i) = \sum_{m=1}^K \pi(m|k) e^{-\frac{1}{2} u_0 (x - \mu_m)' \Gamma^{-1} (x - \mu_m)} e^{\alpha t(x - \mu_m)' \Gamma^{-1} (x_i - \mu_k)}. \quad (31)$$

The conditional independence case corresponds to $\alpha = \beta = \gamma = 0$, $u_0 = 1$, and leads to

$$T_k(x|x_i) = \sum_{m=1}^K \pi(m|k) e^{-\frac{1}{2} (x - \mu_m)' \Gamma^{-1} (x - \mu_m)},$$

which is independent of x_i and proportional to $\sum_{m=1}^K \pi(m|k) f_m(y)$, compare Hjort & Mohn (1984, formula (17)).

The classification procedure (3) with $P_i(k|\Delta_i)$ given by (30) seems to be the simplest one among those that take both types of spatial dependence into account: in the distribution of classes via the transition probabilities $\pi(m|k)$, and between neighbouring spectral measurements, via α, β, γ .

Remark. The class density $f_k(X_i)$ appears in the general formula (23) and in specifications 1, 2, 3, i.e. formulae (24), (26), (30). Since X_i is $N(\mu_{C_i}, \Gamma)$ in the model under study,

$$f_k(x) = (2\pi)^{-d/2} |\Gamma|^{-1/2} e^{-\frac{1}{2} (x - \mu_k)' \Gamma^{-1} (x - \mu_k)}$$

with a common Γ ; i.e. the pooled estimated covariance matrix is the one we should insert. We could however argue that the modelling work done in this section only should be applied to g and h in the adjustment factor $R_k(\Delta_i)$ in the general formula (8), leaving the possibility open of estimating the class density $f_k(X_i)$ by other means, or by other models.

In particular it is often better to assume class dependent covariance matrices, as far as the separate estimation of $f_k(X_i)$ is concerned, that is,

$$f_k(x) = (2\pi)^{-d/2} |\Sigma_k|^{-1/2} e^{-\frac{1}{2} (x-\mu_k)' \Sigma_k^{-1} (x-\mu_k)} \quad (32)$$

can be used as well when applying formulae (24), (26), (30). One could even use say non-parametric density estimation techniques to obtain other estimates of $f_k(X_i)$.

In this way formulae (24), (26), (30) may be looked at as leading to contextual improvements not only of the best linear rule (which has equal covariance matrices), but also of the best quadratic rule (which corresponds to a priori different covariance matrices $\Sigma_1, \dots, \Sigma_K$).

3. ESTIMATION OF PARAMETERS

In Section 2 several contextual classification procedures were proposed, all of the form (3), with neighbourhood data Δ_i as in (5) and $P_i(k|\Delta_i)$ evaluated according to one of formulae (24), (26), (30). Several statistical parameters are involved in these formulae, and estimates are needed before the procedures can be applied in practice.

Assume a training set for each class is available, say

$$Z = \{X_j^{(k)}; j = 1, \dots, n_k, k=1, \dots, K\} \quad (33)$$

with $X_j^{(k)}$ coming from class k . Define the usual class description estimates

$$\mu_k^* = (1/n_k) \sum_{j=1}^{n_k} X_j^{(k)}, \quad (34)$$

$$\Sigma_k^* = (1/n_k) \sum_{j=1}^{n_k} (X_j^{(k)} - \mu_k^*)(X_j^{(k)} - \mu_k^*)'. \quad (35)$$

Following the argumentation in the remark ending Section 2, we might estimate the class density $f_k(x)$ separately, according to either

$$\hat{\Gamma}_k^*(x) = N(\hat{\mu}_k^*, \hat{\Gamma}_k^*)(x)$$

or

$$\hat{\Gamma}_k^*(x) = N(\hat{\mu}_k^*, \hat{\Gamma})(x),$$

where $\hat{\Gamma}^* = \Gamma_{k=1}^K (n_k/n) \hat{\Gamma}_k^*$, $n = \Gamma_{k=1}^K n_k$, is the pooled covariance matrix. $\hat{\Gamma}_k^*$ is no longer an unbiased estimator of Γ_k in view of (12), but its bias is small, and $\hat{\Gamma}_k^*$ is still reasonable to use.

The bias will depend on the location of the pixels in the training set. Assuming that for class k the form is a square of size $m \times m$, it can be shown that

$$E \frac{n_k}{n_{k-1}} \hat{\Gamma}_k^* \approx \Gamma_k - \frac{\theta \Gamma_k}{m^2 - 1} (4\theta + 4\theta^{1/2} + 4\theta^2 - \frac{4\theta + 8\theta^{1/2} + 8\theta^2}{m} + \frac{4\theta^{1/2}}{m^2})$$

where the approximation arises from putting $\theta^x = 0$ for $x > 2$. With for example $\theta = \theta = \frac{1}{2}$ we get

$$E \frac{n_k}{n_{k-1}} \hat{\Gamma}_k^* \approx 0.98 \Gamma_k.$$

Now turn to formulae (24), (26), (30). The prior probabilities $\pi(k)$ may be estimated on the given, unclassified scene by the method in Hjort (1985b), cf. also Hjort & Mohn (1984, section 2). (In fact, the estimates $\hat{\mu}_k^*$, $\hat{\Gamma}_k^*$ may be updated on the way too.) Further involved in (24) are the configuration probability parameters p, q, r . A method producing estimates for these is described in Hjort (1985c). Similarly methods are available in Section 4.4 in Sæbø & al. (1985) and in Hjort (1985b) by which estimates of the transition probabilities $\pi(m|k)$ can be obtained. (Some of the mentioned methods, but not all, require for their use separate sampling of pixel crosses from ground truth.)

What is left, then, is the important problem of estimating α, β, γ defined in (13). Possessing such estimates, other parameters like u, v, w in (17), $t = u + 2v + w$, $u_0 = |S|^{-1/4}$ can automatically also be estimated.

The training set (33) will usually consist of homogeneous groups of pixels, chosen so that a group lies safely within a given class, in particular, a lot of pixel crosses with known classes become available. Assume generally that m_k such crosses, no. i having observations $(X_i^{(k)}, X_{iN}^{(k)}, X_{iE}^{(k)}, X_{iS}^{(k)}, X_{iW}^{(k)})$ are available from class k . (Some of them may overlap.)

Define

$$A^{(k)}(\mu_k) = \frac{1}{m_k} \sum_{i=4}^{m_k} \frac{1}{4} \{ (X_{iN}^{(k)} - \mu_k)(X_i^{(k)} - \mu_k)' + (X_{iE}^{(k)} - \mu_k)(X_i^{(k)} - \mu_k)' + (X_{iS}^{(k)} - \mu_k)(X_i^{(k)} - \mu_k)' + (X_{iW}^{(k)} - \mu_k)(X_i^{(k)} - \mu_k)' \} \quad (36)$$

$$B^{(k)}(\mu_k) = \frac{1}{m_k} \sum_{i=1}^{m_k} \frac{1}{4} \{ (X_{iN}^{(k)} - \mu_k)(X_{iE}^{(k)} - \mu_k)' + \dots + (X_{iW}^{(k)} - \mu_k)(X_{iN}^{(k)} - \mu_k)' \} \quad (37)$$

$$C^{(k)}(\mu_k) = \frac{1}{m_k} \sum_{i=1}^{m_k} \frac{1}{2} \{ (X_{iN}^{(k)} - \mu_k)(X_{iS}^{(k)} - \mu_k)' + (X_{iE}^{(k)} - \mu_k)(X_{iW}^{(k)} - \mu_k)' \}. \quad (38)$$

It should be noted that these matrices are not symmetric. However, using (12) and (13) it easily follows that

$$EA^{(k)}(\mu_k) = \alpha I, \quad EB^{(k)}(\mu_k) = \beta I, \quad EC^{(k)}(\mu_k) = \gamma I.$$

Substituting estimators for the μ_k 's, we should obtain reasonable estimators for αI , βI and γI . The natural choice is to use μ_k^* , defined by (34). However, the expectations of the corresponding estimators will depend on the locations of the pixel crosses, i.e. of the geometrical forms of the training areas, and only general formulae of the expectations can be found. Below we therefore consider an alternative estimator. The expressions derived below for biases and variance, should be conservative approximations to what would result by using μ_k^* instead.

To obtain explicit formulae we will consider the situation where the pixel crosses are chosen so far apart that their spectral vectors can be considered uncorrelated. The central pixels of the crosses can for example be chosen as every 4th pixel in the training set. As estimator for μ_k we will use the mean of the vectors from only the crosses belonging to class k:

$$\bar{X}^{(k)} = \frac{1}{5m_k} \sum_{i=1}^{m_k} (X_i^{(k)} + X_{iN}^{(k)} + X_{iE}^{(k)} + X_{iS}^{(k)} + X_{iW}^{(k)}).$$

The corresponding estimators $A^{(k)}(\bar{X}^{(k)})$, $B^{(k)}(\bar{X}^{(k)})$, $C^{(k)}(\bar{X}^{(k)})$ will just be denoted $A^{(k)}$, $B^{(k)}$ and $C^{(k)}$.

Let us consider $EA^{(k)}$. Introducing

$$U_i = X_i^{(k)} - \mu_k, \quad U_{iN} = X_{iN}^{(k)} - \mu_k, \text{ etc.,}$$

we find for the first term

$$\begin{aligned} E(X_{iN}^{(k)} - \bar{X}^{(k)})(X_i^{(k)} - \bar{X}^{(k)})' &= E U_{iN} U_i' - \frac{1}{5m_k} \sum_{j=1}^{m_k} E(U_j + U_{jN} + U_{jE} + U_{jS} + U_{jW}) U_i' + \\ &\quad - \frac{1}{5m_k} \sum_{j=1}^{m_k} E U_{iN} (U_j + U_{jN} + U_{jE} + U_{jS} + U_{jW})' \\ &+ \frac{1}{25m_k^2} \sum_{i=1}^{m_k} \sum_{j=1}^{m_k} E(U_i + U_{iN} + U_{iE} + U_{iS} + U_{iW})(U_j + U_{jN} + U_{jE} + U_{jS} + U_{jW})' \\ &= \alpha I - \frac{1}{5m_k} (1+4\alpha) I - \frac{1}{5m_k} (1 + \alpha + 2\beta + \gamma) I + \\ &\quad + \frac{m_k}{25m_k^2} (5 + 4\alpha + 4(\alpha + 2\beta + \gamma)) I \\ &= \alpha I - \frac{1}{5m_k} (1 + \frac{17}{5} \alpha + \frac{2}{5} \beta + \frac{1}{5} \gamma) I. \end{aligned}$$

We get the same expression for the other 3 terms. Thus

$$EA^{(k)} = \alpha \bar{\Gamma} - \frac{1}{5m_k} \left(1 + \frac{17}{5} \alpha + \frac{2}{5} \beta + \frac{1}{5} \gamma\right) \bar{\Gamma}. \quad (39)$$

As mentioned in the remark at the end of Section 2 it is more realistic to assume the covariance matrices to be class dependent. In that case we should write Γ_k instead of Γ in (39). Considering the weighted mean of the $A^{(k)}$ in (39). Considering the weighted mean of the $A^{(k)}$'s,

$$A = \frac{1}{m} \sum_{k=1}^K m_k A^{(k)}$$

we find

$$EA = \alpha \bar{\Gamma} - \frac{K}{5m} \left(1 + \frac{17}{5} \alpha + \frac{2}{5} \beta + \frac{1}{5} \gamma\right) \bar{\Gamma}$$

where

$$\begin{aligned} \bar{\Gamma} &= \frac{1}{m} (m_1 \Gamma_1 + \dots + m_K \Gamma_K) \\ \tilde{\Gamma} &= \frac{1}{K} (\Gamma_1 + \dots + \Gamma_K). \end{aligned} \quad (40)$$

In the same way as we obtained (39) we get

$$EB^{(k)} = \beta \Gamma - \frac{1}{5m_k} \left(1 + \frac{2}{5} \alpha + \frac{12}{5} \beta + \frac{6}{5} \gamma\right) \Gamma \quad (41)$$

$$EC^{(k)} = \gamma \Gamma - \frac{1}{5m_k} \left(1 + \frac{2}{5} \alpha + \frac{12}{5} \beta + \frac{6}{5} \gamma\right) \Gamma. \quad (42)$$

We now turn to the covariance matrix of $A^{(k)}$. The calculations become fairly involved, and we therefore simplify matters by considering just the case where m_k is big. Then $\bar{X}^{(k)}$ is close to μ_k , and in the limit, as $m_k \rightarrow \infty$, the covariance matrix of $A^{(k)}$ is equal to the covariance matrix of $A^{(k)}(\mu_k)$, given in (36). We shall find this covariance matrix. Introduce, as above

$$U_i = X_i^{(k)} - \mu_k, \quad U_{iN} = X_{iN}^{(k)} - \mu_k, \quad U_{iE} = X_{iE}^{(k)} - \mu_k, \dots$$

Then

$$A^{(k)}(\mu_k) = \frac{1}{4m_k} \sum_{i=1}^{m_k} (U_{iN}U_i' + U_{iE}U_i' + U_{iS}U_i' + U_{iW}U_i')$$

Let us also introduce the matrices

$$Y_i = U_{iN}U_i' + U_{iE}U_i' + U_{iS}U_i' + U_{iW}U_i', \quad i = 1, \dots, m_k. \quad (43)$$

Then $A^{(k)} = \frac{1}{4m_k} \sum_{i=1}^{m_k} Y_i$. The Y_i 's are identically distributed with $EY_i = 4\alpha I$ and according to our assumption, they are uncorrelated. Hence

$$\text{VAR}(A^{(k)}) = \frac{1}{16m_k^2} \sum_{i=1}^{m_k} \text{VAR}(Y_i) = \frac{1}{16m_k} \text{VAR}(Y_1),$$

(say).

Consider the covariance between two arbitrary elements of Y_1 ; y_{pq} and y_{rs} , say:

$$\text{cov}(y_{pq}, y_{rs}) = Ey_{pq}y_{rs} - Ey_{pq} Ey_{rs}.$$

Because $EY_1 = 4\alpha I$, this equals

$$\text{cov}(y_{pq}, y_{rs}) = Ey_{pq}y_{rs} - 16\alpha^2 \sigma_{pq} \sigma_{rs}.$$

We introduce the vectors U_{iN} , U_{iE} , U_{iS} , U_{iW} , U_i again, using (43). The joint distribution of these vectors are the $5d$ -dimensional multinormal, with expectation zero and covariance matrix

$$\begin{bmatrix} I & \beta I & \gamma I & \beta I & \alpha I \\ & I & \beta I & \gamma I & \alpha I \\ & & I & \beta I & \alpha I \\ & & & I & \alpha I \\ & & & & I \end{bmatrix} \quad (44)$$

Because we don't need the index i any longer, the components of the U -vectors are denoted

$$u_j^N, u_j^E, u_j^S, u_j^W, u_j, \quad j = 1, \dots, d.$$

Then

$$y_{pq} = u_p^N u_q + u_p^E u_q + u_p^S u_q + u_p^W u_q$$

$$y_{rs} = u_r^N u_s + u_r^E u_s + u_r^S u_s + u_r^W u_s$$

and

$$E y_{pq} y_{rs} = E(u_p^N u_q + u_p^E u_q + u_p^S u_q + u_p^W u_q)(u_r^N u_s + u_r^E u_s + u_r^S u_s + u_r^W u_s).$$

The expectation of these 16 terms will only attain three different values, namely

$$E u_p^N u_r^N u_q u_s, \quad E u_p^N u_r^E u_q u_s, \quad E u_p^N u_r^S u_q u_s,$$

depending on the locations in the cross of the factors. There are 4 terms of the first type, 8 of the second and 4 of the third. To find the three values we need the formula for the 4th order moment in the multinormal distribution. Assume then $Z \sim N_p(0, \Lambda)$. Then

$$E Z_i Z_j Z_k Z_l = \lambda_{ij} \lambda_{kl} + \lambda_{ik} \lambda_{jl} + \lambda_{il} \lambda_{jk},$$

cf. Mardia, Kent & Bibby (1979, p.95). In our case Λ is given by (44) and we find

$$E u_p^N u_r^N u_q u_s = \sigma_{pr} \sigma_{qs} + \alpha^2 \sigma_{pq} \sigma_{rs} + \alpha^2 \sigma_{ps} \sigma_{qr}$$

$$E u_p^N u_r^E u_q u_s = \beta \sigma_{pr} \sigma_{qs} + \alpha^2 \sigma_{pq} \sigma_{rs} + \alpha^2 \sigma_{ps} \sigma_{qr}$$

$$E u_p^N u_r^S u_q u_s = \gamma \sigma_{pr} \sigma_{qs} + \alpha^2 \sigma_{pq} \sigma_{rs} + \alpha^2 \sigma_{ps} \sigma_{qr}.$$

Thus

$$\begin{aligned}
 \text{cov}(y_{pq}, y_{rs}) &= 4 \left[(\sigma_{pr} \sigma_{qs} + \alpha^2 (\sigma_{pq} \sigma_{rs} + \sigma_{ps} \sigma_{qr})) \right] \\
 &\quad + 8 \left[\beta \sigma_{pr} \sigma_{qs} + \alpha^2 (\sigma_{pq} \sigma_{rs} + \sigma_{ps} \sigma_{qr}) \right] \\
 &\quad + 4 \left[\gamma \sigma_{pr} \sigma_{qs} + \alpha^2 (\sigma_{pq} \sigma_{rs} + \sigma_{ps} \sigma_{qr}) \right] \\
 &\quad - 16\alpha^2 \sigma_{pq} \sigma_{rs} \\
 &= 4 \left[(1+2\beta+\gamma) \sigma_{pr} \sigma_{qs} + 4\alpha^2 \sigma_{ps} \sigma_{qr} \right].
 \end{aligned}$$

Denoting element (p, q) in $A^{(k)}(\mu_k)$ by $a_{pq}^{(k)}(\mu_k)$ we then have

$$\text{cov}(a_{pq}^{(k)}(\mu_k), a_{rs}^{(k)}(\mu_k)) = \frac{1}{4m_k} \left[(1+2\beta+\gamma) \sigma_{pr} \sigma_{qs} + 4\alpha^2 \sigma_{ps} \sigma_{qr} \right]. \quad (45)$$

For large values of m_k , the covariances between the elements of $A^{(k)}$ are approximately given by this formula. It also follows that

$$\sqrt{m_k}(A^{(k)} - \alpha \Gamma_k) \rightarrow N_{d^2}(0, \Lambda);$$

$$\text{where element } ((p, q), (rs)) \text{ of } \Lambda \text{ is } \frac{(1+2\beta+\gamma) \sigma_{pr} \sigma_{qs} + 4\alpha^2 \sigma_{ps} \sigma_{qr}}{4} \quad (46)$$

(Here σ_{pr}, σ_{qs} etc. are elements of Γ_k ; however, we have suppressed index k .)

The estimation of α, β and γ is considered under the assumptions that the m_k 's are large and that the spectral vectors from different pixel crosses are uncorrelated. Then $EA^{(k)} \approx \alpha \Gamma_k$, or equivalently

$$E a_{pq}^{(k)} \approx \alpha \sigma_{pq}^{(k)}, \text{ for all } p, q \in \{1, \dots, d\}.$$

Thus an estimator of α stemming from training data from class k can be based on a linear combination of the ratios $a_{pq}^{(k)} / \sigma_{pq}^{*(k)}$, where $\{\sigma_{pq}^{*(k)}\} = \tilde{I}_k^*$, given by (35). An appraisal of which linear combination to use can be based on the variances and covariances of the elements of $A^{(k)}$, given by (46).

Let us put

$$\lambda = 1 + 2\beta + \gamma, \quad \sigma_{pq} = \frac{\sigma_{pq}}{\sqrt{\sigma_{pp}\sigma_{qq}}}, \quad p=1, \dots, d, q=1, \dots, p$$

Then from (46)

$$\text{var} \frac{a_{pp}}{\sigma_{pp}} = \frac{\lambda + 4\alpha^2}{4m}, \quad \text{var} \frac{a_{pq}}{\sigma_{pq}} = \frac{1}{4m} \left(\frac{\lambda}{2} + 4\alpha^2 \right), \quad p \neq q,$$

showing that the variances of the diagonal elements of $A^{(k)}$ are less than the off-diagonal elements. The covariance between two diagonal elements is

$$\text{cov} \left(\frac{a_{pp}}{\sigma_{pp}}, \frac{a_{rr}}{\sigma_{rr}} \right) = \sigma_{pr}^2 \frac{\lambda + 4\alpha^2}{4m},$$

which is not so easy to compare with the covariance between the off-diagonal elements. For the special case

$$\text{cov} \left(\frac{a_{pq}}{\sigma_{pq}}, \frac{a_{qp}}{\sigma_{qp}} \right) = \frac{1}{4m} \left(\lambda + \frac{4\alpha^2}{2} \right)$$

the diagonal covariance is clearly smaller. Considering two elements from the same row we find

$$\text{cov} \left(\frac{a_{pq}}{\sigma_{pq}}, \frac{a_{ps}}{\sigma_{ps}} \right) = \frac{1}{4m} \left(\frac{\sigma_{qs}}{\sigma_{pq}\sigma_{ps}} \lambda + 4\alpha^2 \right).$$

This will often be larger than the diagonal terms.

It seems reasonable, therefore, to use just the diagonal terms when estimating α . The unweighted mean is

$$\alpha^*(k) = \frac{1}{d} \sum_{p=1}^d \frac{a_{pp}^{(k)}}{\alpha_{pp}^*(k)} \quad (47)$$

Using the above formula we find, when m_k is large

$$\text{var} \alpha^*(k) \approx \frac{1+4\alpha^2+2\beta+\gamma}{4d^2 m_k} (d + 2 \sum_{r>p} \sigma_{pr}^2) \quad (48)$$

As an example consider the case with $d = 3$ channels. From a simulated SPOT scene of As, Norway, a typical covariance matrix may look like

$$\Sigma = \begin{pmatrix} 10 & 10 & 10 \\ 10 & 20 & 10 \\ 10 & 10 & 100 \end{pmatrix}$$

For the class "arable land" we have found $\alpha^* = 0.76$, $\beta^* = 0.73$, $\gamma^* = 0.71$. Formula (48) gives

$$\text{var} \alpha^*(k) \approx \frac{0.65}{m_k}$$

When $m_k = 100$ this corresponds to a standard deviation of 0.08. For class "blossoming rape seed" we found $\alpha^* = 0.33$, $\beta^* = 0.40$, $\gamma^* = 0.26$, which, with $m_k = 100$, gives a standard deviation of approximately 0.06.

In order to estimate β and γ we proceed in the same way, obtaining estimates

$$\beta^*(k) = \frac{1}{d} \sum_{p=1}^d \frac{b_{pp}^{(k)}}{\alpha_{pp}^*(k)}, \quad \gamma^*(k) = \frac{1}{d} \sum_{p=1}^d \frac{c_{pp}^{(k)}}{\alpha_{pp}^*(k)}, \quad (49)$$

based on training data from the k^{th} class. The variances can be evaluated as above.

Finally, we estimate the correlation parameter ρ and the noise-parameter θ from the definitions (13):

$$\alpha = \theta \rho, \quad \beta = \theta \rho^{\sqrt{2}}, \quad \gamma = \theta \rho^2.$$

It follows that

$$\rho = \left(\frac{\beta}{\alpha} \right)^{\sqrt{2}+1}, \quad \theta = \frac{\alpha^{\sqrt{2}+2}}{\beta^{\sqrt{2}+1}}. \quad (50)$$

Hence, based on training data from the k^{th} class we use the estimators

$$\hat{\rho}^{*(k)} = \left(\frac{\hat{\beta}^{*(k)}}{\hat{\alpha}^{*(k)}} \right)^{\sqrt{2}+1}, \quad \hat{\theta}^{*(k)} = \frac{\hat{\alpha}^{*(k)}}{\hat{\rho}^{*(k)}}.$$

Using data from the above example, we get for "arable land"

$$\hat{\rho}^{*(k)} = \left(\frac{0.73}{0.76} \right)^{\sqrt{2}+1} = 0.94, \quad \hat{\theta}^{*(k)} = \frac{0.76}{0.94} = 0.81$$

and for "blossoming rape seed" $\hat{\rho}^{*(k)} = 0.88$, $\hat{\theta}^{*(k)} = \frac{0.48}{0.88} = 0.49$.

The variances of the estimators $\hat{\rho}^{*(k)}$ and $\hat{\theta}^{*(k)}$ are more difficult to evaluate, because we need the covariance between $\hat{\alpha}^{*(k)}$ and $\hat{\beta}^{*(k)}$. This problem is not studied in this paper.

The estimates obtained for each class have to be combined to one (ρ, θ) -estimate. There are two alternatives: either first to combine the $\hat{\alpha}^{*(k)}$'s, $\hat{\beta}^{*(k)}$'s and $\hat{\gamma}^{*(k)}$'s and then use (50) to obtain $(\hat{\rho}^*, \hat{\theta}^*)$; or to combine the $\hat{\rho}^{*(k)}$'s and $\hat{\theta}^{*(k)}$'s directly. In both cases we combine independent estimators. The optimal weights are then proportional to the inverses of the variances. We see from (48) that the variance of $\hat{\alpha}^{*(k)}$ has the structure τ/m_k , where τ is independent of k . This implies that the weight for class k should equal m_k , the number of pixel crosses. As mentioned above, we have not found the variances of $\hat{\rho}^{*(k)}$ and $\hat{\theta}^{*(k)}$. However, it seems reasonable to use the same weight as for the $\hat{\alpha}^{*(k)}$'s, thus we let

$$\hat{\rho}^* = \frac{1}{m} \sum_k m_k \hat{\rho}^{*(k)}, \quad \hat{\theta}^* = \frac{1}{m} \sum_k m_k \hat{\theta}^{*(k)}.$$

The pixel crosses on which the estimates are based are chosen by picking out every r -th pixel (say) in the training area from the particular class. These pixels serve as the central pixels of the crosses. All pixels of the crosses belong to the class. After obtaining the estimates, the grid of central pixels is moved one step, and a new set of estimates are obtained. The averages of these constitute our final estimate.

An alternative approach to the estimation procedure described above is to use maximum likelihood estimation. Under the assumption that the pixel crosses are independent we can write down the joint distribution of the spectral vectors from the crosses and maximise it with respect to the unknown parameters. The maximum likelihood method is used in Mardia (1984) in a similar problem. We hope to return to this in a forthcoming paper.

4. A MODEL WITH CLASS-DEPENDENT COVARIANCE MATRICES

In this section we consider a slight generalisation of the model. Previously we assumed the same covariance matrix for each class (however, by the Remark at the end of Section 2 we partly circumvented this restriction). Now we allow the covariance matrices to differ from each other by scaling factors.

Specifically we let $X_i = Y_i + \varepsilon_i$, where $Y_1, \dots, Y_M$, $(\varepsilon_1, \dots, \varepsilon_M)$ are independent and where

$$\begin{aligned} Y_i | C_i = k &\sim N(\mu_k, \lambda_k I) \\ \varepsilon_i &\sim N(0, \lambda_0 I) \\ \text{COV}(\varepsilon_i, \varepsilon_j) &= \rho^{|j-i|} \lambda_0 I. \end{aligned} \tag{51}$$

Then

$$\begin{aligned} \text{var } X_i &= (\lambda_k + \lambda_0) I = \tau_k I, \\ \text{say, and } \text{COV}(X_i, X_j) &= \rho^{|j-i|} \lambda_0 I \text{ for } i \neq j. \end{aligned} \tag{52}$$

The model leads to a more awkward formula for the posterior probability (8) than before. It is doubtful whether the combination with Owen-Switzer's mosaic model gives calculations which are feasible in practice. To obtain simplifications when combining with the Markov model requires further approximations.

To find the posterior probability, given by (8) and (9), we start, as previously, by the conditional joint density of the five feature vectors in the cross, given that the classes are a,b,c,d and k. Defining

$$\alpha = \rho \lambda_0, \quad \beta = \rho^2 \lambda_0, \quad \gamma = \rho^2 \lambda_0$$

the covariance matrix is given by

$$\begin{bmatrix} \tau_a & \beta & \gamma & \beta & \alpha \\ & \tau_b & \beta & \gamma & \alpha \\ & & \tau_c & \beta & \alpha \\ & & & \tau_d & \alpha \\ & & & & \tau_k \end{bmatrix} \otimes I$$

where the τ 's are defined by (52). The conditional distribution of X_{iN} , X_{iE} , X_{iS} , X_{iW} , given X_i (and the classes) is again the 4d-dimensional normal with expectation

$$\begin{bmatrix} \mu_a + \alpha/\tau_k (x_i - \mu_k) \\ \cdot \\ \cdot \\ \cdot \\ \mu_d + \alpha/\tau_k (x_i - \mu_k) \end{bmatrix}$$

and covariance matrix $S_k(a,b,c,d) \otimes I$, where

$$S_k(a,b,c,d) = \begin{bmatrix} \tau_a - \alpha^2/\tau_k & \beta - \alpha^2/\tau_k & \gamma - \alpha^2/\tau_k & \beta - \alpha^2/\tau_k \\ & \tau_b - \alpha^2/\tau_k & \beta - \alpha^2/\tau_k & \gamma - \alpha^2/\tau_k \\ & & \tau_c - \alpha^2/\tau_k & \beta - \alpha^2/\tau_k \\ & & & \tau_d - \alpha^2/\tau_k \end{bmatrix} \quad (53)$$

The difficulty now is that this matrix depends on the classes a, b, c, d . All elements in its inverse will depend on the classes. We denote the inverse by

$$S_k(a, b, c, d)^{-1} = \{s_k^{pq}(a, b, c, d)\}$$

and let

$$\sigma_k^q(a, b, c, d) = \sum_{p=1}^4 s_k^{pq}(a, b, c, d), \quad q = 1, 2, 3, 4$$

$$\sigma_k(a, b, c, d) = \sum_{q=1}^4 \sigma_k^q(a, b, c, d)$$

be the column sums and total sum of the elements in the inverse. Then the posterior probability can be written

$$P_i(k|\Delta_i) = \text{const } \pi(k) f_k(X_i)$$

$$\tau_{a,b,c,d} g(a, b, c, d|k) |S_k(a, b, c, d)|^{-\frac{d}{2}} e^{-\frac{1}{2} Q_k(a, b, c, d)}, \quad (54)$$

where

$$Q_k(a, b, c, d) = y_i' (S_k(a, b, c, d)^{-1} \otimes \Gamma^{-1}) y_i + \frac{\alpha^2}{\tau_k} \sigma_k(a, b, c, d) (X_i - \mu_k)' \Gamma^{-1} (X_i - \mu_k) - 2 \frac{\alpha}{\tau_k} (X_i - \mu_k)' \Gamma^{-1} [\sigma_k^1(a, b, c, d) (X_{iN} - \mu_a) + \dots + \sigma_k^4(a, b, c, d) (X_{iW} - \mu_d)] \quad (55)$$

Here y_i is given by (20).

Compared to (23) the main drawback of (54) is that we must calculate the inverse and the determinant of the 4×4 matrix $S_k(a, b, c, d)$ for each term in the sum. In order to use (54) in practice it is necessary to combine it with g -functions which lead to significant simplifications.

Specification 1

Using the generalisation of the Owen Switzer model we find

$$P_i(k|\Delta_i) = \text{const } \pi(k) f_k(X_i)$$

$$\{[p+(q+r)\pi(k)] |S_k(k,k,k,k)|^{-\frac{d}{2}} e^{-\frac{1}{2} Q_k(k,k,k,k)} +$$

$$+ \frac{q}{4} \sum_{m \neq k} \pi(m) D_2(k,m) \left(e^{-\frac{d}{2}} e^{-\frac{1}{2} Q_k(k,k,m,m)} + \dots + e^{-\frac{1}{2} Q_k(k,m,m,k)} \right)$$

$$+ \frac{r}{4} \sum_{m \neq k} \pi(m) D_1(k,m) \left(e^{-\frac{d}{2}} e^{-\frac{1}{2} Q_k(k,k,m,k)} + \dots + e^{-\frac{1}{2} Q_k(k,m,k,k)} \right),$$

where

$$D_2(k,m) = |S_k(k,k,m,m)| = \dots = |S_k(k,m,m,k)|$$

$$D_1(k,m) = |S_k(k,k,m,k)| = \dots = |S_k(k,k,m,k)|.$$

From (55) we see that the first term of Q_k is a quadratic form in 4d dimensions with a matrix which has to be calculated for each combination of classes. The two other terms could as well be calculated in this space. Thus

$$Q_k(a,b,c,d) = [X_{iN}' - \mu_a' - v_k', \dots, X_{iW}' - \mu_d' - v_k'] (S_k(a,b,c,d)^{-1} \otimes I^{-1}) \begin{bmatrix} X_{iN}' - \mu_a' - v_k' \\ \vdots \\ X_{iW}' - \mu_d' - v_k' \end{bmatrix}$$

where $v_k = \frac{\alpha^2}{\tau_k} (X_i - \mu_k)$. The formula may still give too heavy computations to be of any practical use. A significant simplification in the calculations is obtained when we turn to

Specification 2

This model corresponds to Switzers (1980) "local spatial continuity" and is obtained by putting $p = 1$, $q = r = 0$. Then

$$P_i(k|\Delta_i) = \text{const } \pi(k) f_k(X_i) |S_k(k,k,k,k)|^{-\frac{d}{2}} e^{-\frac{1}{2} Q_k(k,k,k,k)}.$$

The matrix $S_k(k,k,k,k)$ has the structure $\begin{bmatrix} a & b & c & b \\ & a & b & c \\ & & a & b \\ & & & a \end{bmatrix}$, with all elements

dependent on k . Then the inverse has the same structure

$$S_k(k,k,k,k)^{-1} = \begin{bmatrix} u & v & w & v \\ & u & v & w \\ & & u & v \\ & & & u \end{bmatrix}.$$

say, giving $\sigma_k^q(k,k,k,k) = u(k) + 2v(k) + w(k) = t(k)$, $\sigma_k(k,k,k,k) = 4t(k)$.

Then

$$P_i(k|\Delta_i) = \text{const } \pi(k) f_k(X_i) |S_k|^{-\frac{d}{2}} R_i(k)$$

$$\exp\left\{-\frac{2\alpha^2 t(k)}{\tau_k} (X_i - \mu_k)' \Gamma^{-1} (X_i - \mu_k) + 4 \frac{\alpha t(k)}{\tau_k} (X_i - \mu_k)' \Gamma^{-1} (\bar{X}_i - \mu_k)\right\},$$

where $\bar{X}_i = \frac{1}{4} (X_{iN} + X_{iE} + X_{iS} + X_{iW})$ and

$$R_i(k) = \exp\left\{-\frac{1}{2} [X'_{iN} - \mu'_k, \dots, X'_{iW} - \mu'_k] \begin{bmatrix} u(k) & v(k) & w(k) & v(k) \\ & u(k) & v(k) & w(k) \\ & & u(k) & v(k) \\ & & & u(k) \end{bmatrix} \otimes \Gamma^{-1} \begin{bmatrix} X_{iN} - \mu_k \\ \vdots \\ X_{iW} - \mu_k \end{bmatrix}\right\}.$$

Specification 3

In order to obtain a simplification when using the Markov model (28)

$$g(a,b,c,d|k) = \pi(a|k) \pi(b|k) \pi(c|k) \pi(d|k),$$

the rest of the terms in (54) must be written as a product of factors, each depending on only one of the summation indexes a, b, c, d . In the previous case this was obtained by approximating S^{-1} by the diagonal $u_0 I_4$. Now S^{-1} depends on k, a, b, c and d . Approximating with $u_k(a, b, c, d) I_4$ in such a way that

$$|u_k(a, b, c, d) I_4| = |S_k(a, b, c, d)^{-1}|,$$

we get

$$u_k(a, b, c, d) = |S_k(a, b, c, d)|^{-\frac{1}{4}}.$$

The approximation implies that

$$\sigma_k^q(a, b, c, d) u_k(a, b, c, d), \quad \sigma_k(a, b, c, d) 4u_k(a, b, c, d)$$

and the exponent (55) can be approximated by

$$\begin{aligned} Q_k(a, b, c, d) &= \frac{u_k(a, b, c, d)}{2} [(X_{iN} - \mu_a)' \Gamma^{-1} (X_{iN} - \mu_a) + \dots + (X_{iW} - \mu_d)' \Gamma^{-1} (X_{iW} - \mu_d)] \\ &+ 4 \frac{\alpha^2 u_k(a, b, c, d)}{\tau_k} (X_i - \mu_k)' \Gamma^{-1} (X_i - \mu_k) \\ &- 2 \frac{\alpha u_k(a, b, c, d)}{\tau_k} (X_i - \mu_k)' \Gamma^{-1} (X_{iN} - \mu_a + \dots + X_{iW} - \mu_d). \end{aligned}$$

Then

$$P_i(k|\Delta_i) = \text{const } \pi(k) f_k(X_i)$$

$$a, b, c, d \pi(a|k) \pi(b|k) \pi(c|k) \pi(d|k) u_k(a, b, c, d)^{2d} e^{-\frac{1}{2} Q_k(a, b, c, d)}.$$

To proceed further we first factorise $u_k(a,b,c,d)$ by approximating the determinant of (53) with the product of the diagonal terms:

$$u_k(a,b,c,d)^{-4} = |S_k(a,b,c,d)| \approx (\tau_a - \frac{\alpha^2}{\tau_k}) (\tau_b - \frac{\alpha^2}{\tau_k}) (\tau_c - \frac{\alpha^2}{\tau_k}) (\tau_d - \frac{\alpha^2}{\tau_k}).$$

In the exponent we need a value independent of the summation indexes. Using the geometric mean of the τ_k 's

$$\bar{\tau}_G = \sqrt[K]{\tau_1 \tau_2 \dots \tau_K},$$

we define u_k by

$$u_k^{-4} = (\bar{\tau}_G - \frac{\alpha^2}{\tau_k})^4, \quad u_k = 1/(\bar{\tau}_G - \frac{\alpha^2}{\tau_k})$$

and let $u_k(a,b,c,d) \approx u_k$. We then obtain

$$P_i(k|\Delta_i) \approx \text{const } \pi(k) f_k(X_i) e^{-\frac{2\alpha^2 u_k}{\tau_k} (X_i - \mu_k)' \Gamma^{-1} (X_i - \mu_k)} \quad (56)$$

$$T_k(X_{iN}|X_i) T_k(X_{iE}|X_i) T_k(X_{iS}|X_i) T_k(X_{iW}|X_i),$$

where

$$T_k(x|x_i) = \prod_{m=1}^K \pi(m|k) (\tau_m - \frac{\alpha^2}{\tau_k}) e^{-\frac{u_k}{2} (x - \mu_m)' \Gamma^{-1} (x - \mu_m)} \\ \propto \frac{u_k}{\tau_k} (x_i - \mu_k)' \Gamma^{-1} (x - \mu_m) e$$

If we assume the spectral vectors conditionally independent, corresponding to $\alpha = \beta = \gamma = 0$, the formula gives

$$P_i(k|\Delta_i) \approx \text{const } \pi(k) f_k(k) T_k(X_{iN}) T_k(X_{iE}) T_k(X_{iS}) T_k(X_{iW}),$$

where

$$T_k(x) = \sum_{m=1}^K \pi(m|k) \tau_m e^{-\frac{1}{2\tau_m} (x - \mu_m)' \Sigma^{-1} (x - \mu_m)}.$$

The parameters $\tau_1, \dots, \tau_K$ must be estimated from training data. The problem has been discussed by Owen & Switzer (1982) and will not be considered in this paper.

REFERENCES

- Anderson, T.W. (1958). Introduction to multivariate analysis. Wiley.
- Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. J.R. Stat. Soc., Series B, 36, p. 192-236.
- Haslett, J. (1983). Maximum likelihood discriminant analysis on the plane using a Markovian model of spatial context. Dep. of stat., Trinity College, Dublin.
- Hjort, N.L. (1985a). Detecting outliers. In Sæbø & al.
- Hjort, N.L. (1985b). Estimating parameters in neighbourhood based classifiers for remotely sensed data, using unclassified vectors. In Sæbø & al.
- Hjort, N.L. (1985c). Neighbourhood based classification of remotely sensed data based on geometric probability models. In Sæbø & al.
- Hjort, N.L. & Mohn, E. (1984). A comparison of some contextual methods in remote sensing classification. Eighteenth International Symposium on Remote Sensing of Environment. CNES, Paris.
- Kittler, J. & Föglein, J. (1984). Contextual classification of multi-spectral pixel data. Image and vision computing 2, p. 13-29.
- Mardia, K.V. (1984). Spatial discrimination and classification maps. Commun. Statist. - Theor. Meth. 13(18), p. 2181-2197.
- Mardia, K.V., Kent, J.T. and Bibby, J.M. (1979). Multivariate Analysis. Academic Press.
- Owen, A. & Switzer, P. (1982). A neighbourhood based classifier for Landsat data, Techn. Report, Dep. of stat., Stanford University.
- Pickard, D.K. (1980). Unilateral Markov fields, Adv. appl. prob. 12, p. 655-671.
- Switzer, P. (1980). Extensions of linear discriminant analysis for statistical classification of remotely sensed satellite imagery. J. Inter. Assoc. Mathl. Geol. 12, p. 367-376.
- Sæbø, H.V., Bråten, K.H., Hjort, N.L., Llewellyn, B., and Mohn, E. (1985). Contextual classification of remotely sensed data: statistical methods and development of a system. Project Report, Norwegian Computing Center, Oslo, Norway.