SOME SPATIAL STATISTICS FOR THE INTERPRETATION OF SATELLITE DATA

BY

PAUL SWITZER

TECHNICAL REPORT NO. 4
JUNE 30, 1983

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

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

SOME SPATIAL STATISTICS FOR THE INTERPRETATION OF SATELLITE DATA

BY

PAUL SWITZER
STANFORD UNIVERSITY

TECHNICAL REPORT NO. 4

JUNE 30, 1983

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

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

SOME SPATIAL STATISTICS FOR THE INTERPRETATION OF SATELLITE DATA

Paul Switzer
Stanford University

ABSTRACT

This paper reviews concepts of spatial pre-smoothing and post-smoothing for improvement of classification maps, particularly of the type derived from gridded data common to satellite data. Examples show the type of improvement to be expected. However, most smoothers are not well behaved near boundaries between classes. Formalization of boundary processes enable one to examine probabilistically the behavior of smoothers near boundaries. Some formulas are given which relate classification error rates to the size of the smoothing window and which provide estimates for the minimum achieveable error rate.

SOME SPATIAL STATISTICS FOR THE INTERPRETATION OF SATELLITE DATA

Paul Switzer
Stanford University

1. INTRODUCTION

There is a large literature on image processing of satellite data. A recent survey of classification procedures appears in Rosenfeld and Kak (1982; Ch. 10). However, algorithms for the classification of surface pixels using multi-dimensional spectral data do not commonly make use of the special geographic nature of the data. Substantial spatial autocorrelation between the data vectors of nearby pixels should be exploited both in the design of classifiers and in the estimation of their error rates. Such correlations may be made part of probability models in several ways leading either to pre-smoothing or post-smoothing modifications of standard classification algorithms. Pre-smoothing refers to spatial filtering of the input data while post-smoothing refers to spatial regularization of class assignments. Smoothers may be expected to improve the classification accuracy in the interior of regions which are class homogeneous and large relative to the size of single pixels. However, smoothers will also blur the boundary between classes, and improvements along or near the boundary are not to be expected. Special attention should be paid to the behavior of classifiers near class boundaries. Class boundaries may be described locally in probabilistic terms. Such descriptions lead to representations for the minimum achievable error rate. We will present some underlying probability theory with its implications for classification methodology together with an example of application to LANDSAT satellite imaging.

2. NOTATION AND STRUCTURE

The m-variate data are denoted by

$$\underset{\sim}{Z}(x) = (Z_1(x), Z_2(x), \ldots, Z_m(x))' \qquad x \in X$$

where $Z_i(x)$ is scaled measurement of emitted or reflected energy in wavelength band i, integrated over a geographic pixel (unit square) centered at $x = (x_1, x_2)$.

The mathematical structure for the pixel data is given in terms of integrated stochastic processes defined at all points of the geographic domain:

$$\underset{\sim}{Z}(x) = \int_{\eta_1} [\underset{\sim}{S}(x+t) + \underset{\sim}{\varepsilon}(x+t)]dt$$

where

$\underset{\sim}{S}(x)$ is the m-variate "signal" at the geographic point x

$\underset{\sim}{\varepsilon}(x)$ is the m-variate "noise" at x

dt is the differential of Lebesgue measure in the plane

η_1 is the unit square pixel centered at the origin and oriented like the data.

The classification structure is a specialization of the model above wherein the signal has only a small number of possible values, i.e.,

$$\underset{\sim}{S}(x) = \underset{\sim}{\mu}(I(x))$$

where $I(x)$ is the class (e.g. rock type) of the surface at point x. $I(x)$ belongs to a finite set; there are only K possible values for $\underset{\sim}{S}$ or $\underset{\sim}{\mu}$. Customarily, the class indicator $I(x)$ is taken to be constant over the area of a single pixel, hence the data model specializes further and becomes

$$\underset{\sim}{Z}(x) = \underset{\sim}{\mu}(I(x)) + \underset{\sim}{\varepsilon}_1(x) \quad \text{where} \quad \underset{\sim}{\varepsilon}_1(x) = \int_{\eta_1} \underset{\sim}{\varepsilon}(x+t)dt \ .$$

The objective is to infer $I(x)$ from $\underset{\sim}{Z}(x)$ for each pixel. Later the consequences of assuming homogeneous pixels, i.e. $I(x)$ constant within each pixel, will be explored.

The noise process $\underset{\sim}{\varepsilon}(x)$ is taken to be a zero-mean stationary spatially correlated multivariate stochastic process. Historically, the signal $\underset{\sim}{S}(x)$ has sometimes been treated as random and sometimes as fixed. Maximum likelihood type classifiers have treated the class indicator $I(x)$ as a fixed function, whereas Bayes classifiers treat the class indicator $I(x)$ as a random spatial process.

"Standard" linear discriminant analysis (LDA) is an ML classifier when $\underset{\sim}{\varepsilon}(x)$ is multivariate Gaussian white noise. Therefore, LDA makes no use of geographic continuity.

"Standard" Bayes classifiers treat $\underset{\sim}{\varepsilon}(x)$ as white noise and $I(x)$ as sampled from a fixed multinomial distribution, independently for each pixel. Therefore, the usual Bayes classifiers also do not make explicit use of geographic continuity. The incorporation of geographic continuity into classification algorithms may be accomplished either by "pre-smoothing" the original data or by "post-smoothing" the resulting classification maps.

3. PRE-SMOOTHING AND POST-SMOOTHING

Pre-smoothing describes data filtering prior to classification. We define a filtered version of the data in terms of a moving average

$$\underset{\sim}{Z}^*(x) = \sum_{\Delta \in \eta} w(\Delta) \cdot Z(x+\Delta)$$

where η is a geographic neighborhood of pixels and $w(\Delta)$ is a weight function depending on the displacement Δ. We then apply standard pointwise estimators to the spatially filtered data $\underset{\sim}{Z}^*$.

(Simplest: If η is a square window of $N \times N$ pixels and $w(\Delta) = \frac{1}{N^2}$.)

The model justification for such moving averages is discussed in Switzer (1980); the basic ingredients are:

i. The class indicator $I(x)$ is constant over the neighborhood η

ii. $\text{cov}\{\underset{\sim}{\varepsilon}(x), \underset{\sim}{\varepsilon}(x+\Delta)\} = \gamma(\Delta) \cdot \Sigma_{m \times m}$

iii. $\underset{\sim}{\varepsilon}(x)$ is Gaussian.

Then simple pre-smoothing by means of a moving average filter, followed by standard LDA, is the maximum likelihood classifier. The weighting function $w(\Delta)$ may be derived from the spatial covariance attenuation function $\gamma(\Delta)$ as described in the abovementioned reference.

Post-smoothing may be described in the following terms. Let $\hat{I}(x)$ be the class assignments resulting form a standard pointwise (non-spatial) classifier. If the plurality of the pixels in a neighborhood of x have been assigned to class k initially then the pixel at x may be reassigned to class k. This will tend to give a smoother classification map with lower error probabilities. (The plurality may be based on distance-weighted voting.)

A somewhat more sophisticated approach has been suggested in Switzer, Kowalik, and Lyon (1981). Consider the class indicator $I(x)$ to be stochastic, i.e. $I(x) = k$ with probability $\pi_k(x)$, $k=1, \ldots, K$, $\sum \pi_k(x) = 1$, for each pixel centered at x. This prior probability vector, $\underset{\sim}{\pi}(x) = (\pi_1(x), \ldots, \pi_k(x))'$, is assumed to vary with x thereby introducing spatial considerations. However, if $\underset{\sim}{\pi}(x)$ is approximately the same for all pixels in a neighborhood of x, then it is possible to estimate $\underset{\sim}{\pi}(x)$ from the data and then use the estimate as input to a "standard" Bayes classifier.

The estimation of $\underset{\sim}{\pi}(x)$ may proceed as follows:

i. Use any naive classifier to obtain preliminary class assignments $\hat{I}(x)$ for each pixel x.

ii. For the training data with known class identities calculate the "confusion" matrix $\underset{\sim}{f}_{K\times K}$ where

f_{ij} = proportion of class j pixels which have been assigned to class i by the naive classifier; $\sum_i f_{ij} = 1$.

iii. For each pixel x calculate the proportion $p_i(x)$ of pixels in a fixed neighborhood η of x which have been assigned to class i by the naive classifier; $\underset{\sim}{p}(x) = (p_1(x), \ldots, p_k(x))'$.

iv. In terms of expected values

$$E[\underset{\sim}{p}(x)] = \underset{\sim}{\pi}'(x)\ E[\underset{\sim}{f}]$$

if the confusion matrix is taken to be geographically constant. Hence, the local a priori class probabilities are estimated by

$$\hat{\underset{\sim}{\pi}}(x) = \underset{\sim}{f}^{-1}\ \underset{\sim}{p}(x)\ .$$

The estimated prior probabilities for each pixel are then used with a standard Bayes classifier.

4. APPLICATION

The pre-smoothing and post-smoothing methods for extracting information from contiguous pixels are compared empirically for a 16 × 25 rectangular test area, located in western Nevada. Each pixel is about an acre in size and for each pixel there are four measurements, consisting of energy fluxes in four wavelength bands as measured by the LANDSAT earth satellite. It is

known that each pixel belongs to one of three major rock types - volcanic, limonitically altered Jurassic quartz monzonite, and Jurassic quartz monzonite (not limonitically altered). Out of 400 pixels 158 are volcanic, 183 are limonitically altered, and 59 are nonlimonitic. Out of these 400 pixels 20, 20, and 14 pixels, respectively, were selected along linear paths to comprise the training data set. The true map of the test area is shown in Figure 1.

The first attempt at categorizing the pixels by the satellite data is standard linear discriminant analysis (LDA), using the mean data vectors of training pixels for each rock type, and a pooled covariance matrix. The map obtained using the naive procedure is shown in Figure 2 and its overall error rate is 39%.

Since the geograhpic regions occupied by each rock type are generally large with respect to the size of a single pixel, smoothing techniques which exploit this spatial homogeneity should improve upon the usual linear discriminant analysis. Pre-smoothing consisted of passing a 3×3 equal weights window over the area. The resulting average data vector replaces the data vector for the center pixel. Using the smoothed data we recalculate the predicted categories for each pixel, with the linear discriminant functions derived from our original analysis. The resulting classification map is shown in Figure 3, and its overall error rate is 25%.

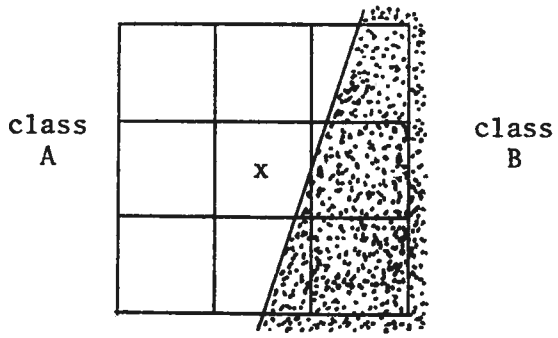
For comparison, the same data were treated to post-smoothing using a 3×3 window. Post-smoothing estimates of the geographically varying prior class of probabilities were incorporated into a Bayes classifier using a conditional multivariate normal model for the distribution of the data vector $\underset{\sim}{Z}$ in each class. The result of this operations was to add $2|\ell n\ \pi_i(x)|$

to the Mahalanobis distances between $\underset{\sim}{Z}$ and the mean vectors for each rock type as determined from the training data. (Negative estimates of π_i can occur; in such cases they are reset to zero and the positive class probabilities are rescaled.) The result of this post-smoothing of the original classification map is shown in Figure 5. It is worth noting that both pre-smoothing and post-smoothing gave similar overall reductions in the error rate although the classification maps are somewhat different. The estimated prior probabilities for the limonitic pixels are shown in Figure 4.

Finally, it is possible to perform both pre-smoothing and post-smoothing to the data. In this case, the initial classification is that based on pre-smoothed data as described above, from which one gets the estimated confusion matrix (f_{ij}). This matrix is then used in the post-smoothing operation based on estimated geographically varying prior class probabilities, as described above. The results are shown in Figure 6 and a substantial further improvement in the classification is noted. Figure 7 shows the estimated classifications of the limonitically altered pixels only and it may be perceived, as expected, that most classification eerors occur near the class boundary. This suggests that an accuracy limit may have been reached.

5. BOUNDARY PROBLEMS

In the preceding sections we have acted as though pixels or neighborhoods of pixels were homogeneous with respect to their class identification. We now consider the implications of class boundaries cutting across pixels.



The point x is the center of a pixel but now $I(x)$ need not be constant over the whole pixel or a neighborhood of pixels. For simplicity of presentation, we consider the case of only two possible classes A, B. Consider the problem of classifying the point x using the averaged data in an $N\times N$ neighborhood η_N centered at x, denoted $\underset{\sim}{Z}_N(x)$. The unit pixel "neighborhood" is η_1. We may write

$$\underset{\sim}{Z}_N(x) = \frac{1}{N^2}\int_{\eta_N} \underset{\sim}{Z}(x+t)dt = \frac{1}{N^2}\int_{\eta_N} [\underset{\sim}{\mu}(I(x+t)) + \underset{\sim}{\varepsilon}(x+t)]dt$$

$$= \lambda_N(x)\cdot\underset{\sim}{\mu}(A) + [1 - \lambda_N(x)]\cdot\underset{\sim}{\mu}(B) + \underset{\sim}{\varepsilon}_N(x)$$

where

$\lambda_N(x)$ = proportion of the area of the $N\times N$ neighborhood of x which belongs to class A

$\underset{\sim}{\varepsilon}_N(x)$ = noise averaged over the $N\times N$ neighborhood of x.

Let $\text{cov}(\underset{\sim}{\varepsilon}_N(x)) = \underset{\sim}{\Sigma}_N$. A common approximation is to take $\underset{\sim}{\Sigma}_N = b_N^2\underset{\sim}{\Sigma}_1$. In this case the LDA classifier is the same for all N, although the classification error rate will, of course, depend on the size of the neighborhood. If $\underset{\sim}{\varepsilon}(x)$ is white noise then $b_N = 1/N$. If $\underset{\sim}{\varepsilon}(x)$ is positively autocorrelated then $b_N \geq 1/N$.

We now examine how the pointwise classification error probability depends on the size of the averaging neighborhood. Suppose we have a linear classifier, i.e.

$$x\mapsto A \quad (x \text{ is assigned class A}) \quad \text{iff} \quad \underset{\sim}{c}'\underset{\sim}{Z}_n(x) \leq L \ .$$

The error probability for a point $x\in A$ is

$$\alpha \equiv 1\text{-}P\{x\mapsto A \,|\, x\in A\}$$

where

$$P\{x \to A \,|\, x \in A\} = P\{\underset{\sim}{c}' \underset{\sim}{Z}_N(x) \leq L \,|\, x \in A\}$$

$$= E_{\lambda_N} P\{\underset{\sim}{c}' \underset{\sim}{Z}_N(x) \leq L \,|\, x \in A,\ \lambda_N\}$$

and E_{λ_N} denotes an average over the distribution of λ_N. If the noise spatial process $\underset{\sim}{\varepsilon}(x)$ is Gaussian, then, given λ_N, $\underset{\sim}{c}' \underset{\sim}{Z}_N(x)$ is a Gaussian random variable with

$$\text{mean} = \lambda_N \underset{\sim}{c}' \underset{\sim}{\mu}(A) + (1-\lambda_N)\, \underset{\sim}{c}' \underset{\sim}{\mu}(B)$$

$$\text{variance} = b_N^2 \underset{\sim}{c}' \underset{\sim\sim}{\Sigma} \underset{\sim}{c} \quad \text{(not depending on } \lambda_N\text{)} .$$

Therefore, for a geographic point x belonging to class A, the conditonal misclassification error rate for fixed λ_N has the form

$$\Phi\{\Delta - \tau \cdot (\lambda_N - \tfrac{1}{2})/b_N\}$$

where Δ, τ do not depend on N, the size of the averaging window. If π_A is the prior probability for class A (which may depend on location x) and if the Bayes classifier is used then $\Delta = [\ell n \frac{1-\pi_A}{\pi_A}]/\tau$ and $\tau^2 = [\underset{\sim}{\mu}(A) - \underset{\sim}{\mu}(B)]' \underset{\sim}{\Sigma}_1^{-1} [\underset{\sim}{\mu}(A) - \underset{\sim}{\mu}(B)]$.

(At this point we may ask when it is better to use a single pixel instead of neighborhood averages. For a fixed value of λ_N, the conditional error rate at x when $x \in A$ is a decreasing function of $(\lambda_N - \frac{1}{2})/b_N$ since Δ and τ do not depend on N. So averaging is advantageous if

$$\frac{(\lambda_N - \tfrac{1}{2})}{b_N} > (\lambda_1 - \tfrac{1}{2}) .$$

For example, suppose $\lambda_1 = \frac{1}{2}$; then averaging is always better because $\lambda_N \geq \frac{1}{2}$ for any N. Now suppose $\lambda_1 = 1$, i.e. the central pixel is all

in class A, then averaging is better when $\lambda_N \geq \frac{1}{2}(1+b_N)$. If the spatial autocorrelation is very strong then b_N is close to unity and averaging makes things worse. If $\underset{\sim}{\varepsilon}(x)$ is an uncorrelated residual process then $b_N = 1/N$; then λ_N is necessarily $\geq \frac{1}{2}(1+1/N)$ when $\lambda_1=1$, so all N×N averages are better than single pixel data in this case.)

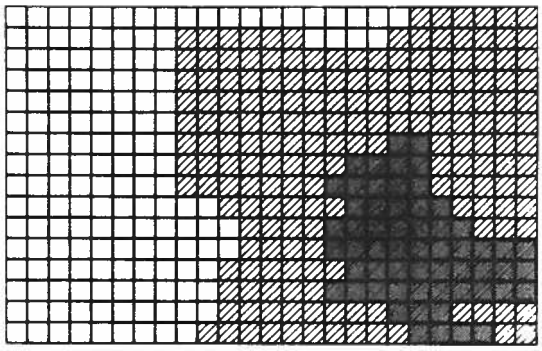


FIG. 1 'TRUE' CLASS MAP (400 Pixels)
west-central Nevada

- Volcanic rock
- Limonitic rock
- Nonlimonitic rock

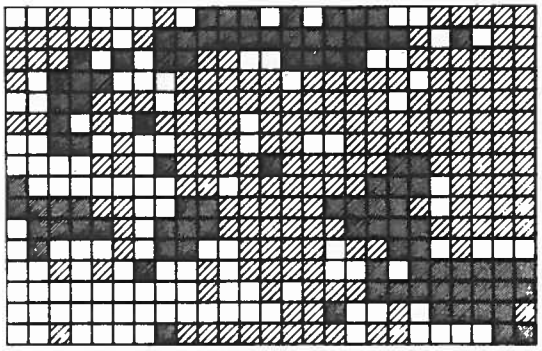


FIG. 2 NAIVE LDA CLASS ASSIGNMENTS
NO SMOOTHING
Error rate ≈ 39%

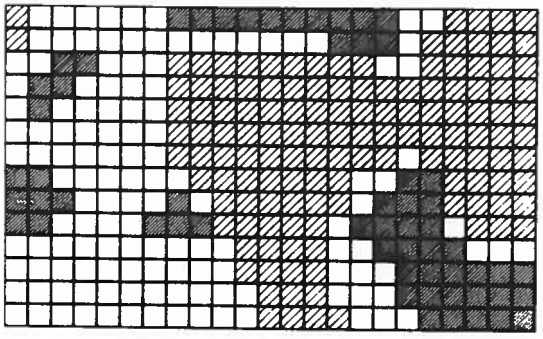


FIG. 3 CLASS ASSIGNMENTS
with PRESMOOTHING (3 x 3 window)
Error rate ≈ 25%

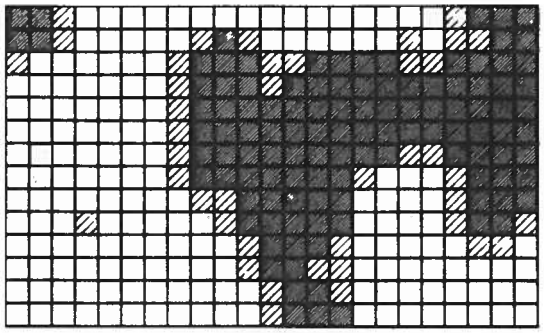


FIG. 4 ESTIMATED a priori PROBABILITY for LIMONITIC
(3 x 3 window applied to LDA assignments)

high probability

medium probability

low probability

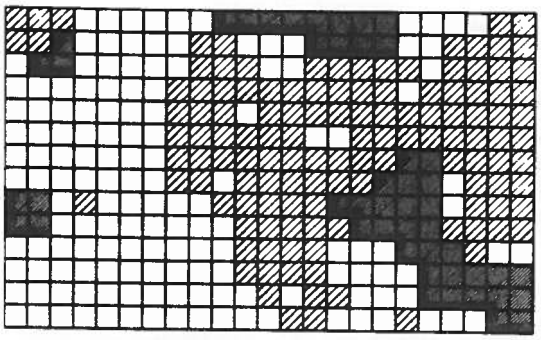


FIG. 5 CLASS ASSIGNMENTS
with post-smoothing (3 x 3 window)
Error rate ≈ 27%

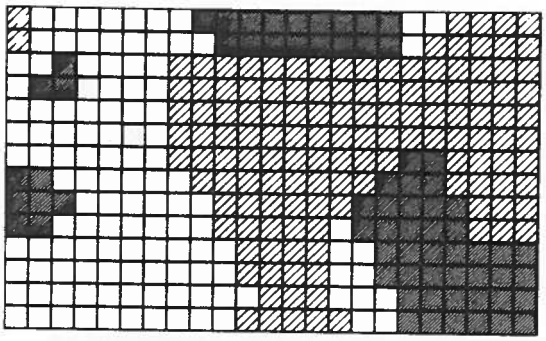


FIG. 6 CLASS ASSIGNMENTS
with pre-smoothing (3 x 3 window)
and post-smoothing (3 x 3 window)
Error rate ≈ 19%

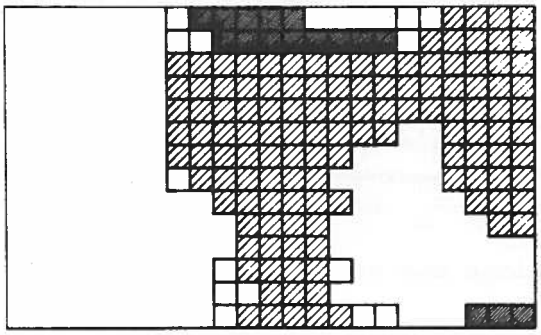


FIG. 7 CLASS ASSIGNMENTS of LIMONITIC PIXELS
with pre-smoothing
and post-smoothing

The conditional error rate for fixed λ_N must now be averaged over the distribution of λ_N, the proportion of the $N \times N$ neighborhood falling into class A. To obtain a distribution for λ_N we use a simplified probability specification for the occurrence of class boundaries. If a boundary occurs within the square neighborhood it is taken to be a line segment generated by the invariant measure. This measure is the product of uniform measures for the slope of the line segment and its distance to the center of the neighborhood. Reference to the invariant measure may be found in Kendall (1963).

We shall further assume that the probability ν that a pixel is intersected by a boundary segment is small and that the probability of two or more segments crossing a pixel is negligible. The invariant measure then implies that the probability of a boundary segment crossing an $N \times N$ neighborhood of pixels is approximately $\nu \cdot N$ if N is not too large, i.e. it is proportional to the linear dimension of the neighborhood.

The class categories on either side of the boundary are taken to be independent random variables, identically distributed each with probabilities π_A, $1-\pi_A$ for this two class example. The independence implies that the same class may occur on both sides of a "boundary" segment. This formulation permits the rate parameter ν to be separated from the class frequencies and is especially convenient for generalizations to more than two possible classes.

The preceding boundary model leads to a calculation of the distribution function, G_N, of the random variable λ_N = proportion of an $N \times N$ square neighborhood belonging to class A, given that the center $x \in A$. The calculation involves the area of the larger portion of a square partitioned by a random line segment $[x \in A \Leftrightarrow \lambda_N \geq \frac{1}{2}]$. A representation for G_n is

$$P\{\lambda_N \leq t \mid x\epsilon A\} \equiv G_N(t) = [1-\pi_A]\cdot\nu\cdot N\cdot J(2-2t), \quad \text{for} \quad \tfrac{1}{2} \leq t < 1 ,$$

where

$$J(u) = \frac{(1-u^2)}{\sqrt{1+u^2}} + \frac{\sqrt{2u}}{2}\int_0^{w(u)} \sqrt{\cos\,\theta}\; d\theta$$

and

$$w(u) = \frac{\pi}{2} - 2\,\tan^{-1} u .$$

The distribution function is discontinuous at $\lambda_N = 1$ which has positive probability. The unconditional error probability α for misclassifying a point $x\epsilon A$ is then given by

$$\alpha = \int_{.5}^{1} \Phi\{\Delta - \tau\cdot(t-\tfrac{1}{2})/b_N\}\; dG_N(t) .$$

The probability specification under which the classification error rate has been calculated may be regarded as a minimal specification incorporating spatial information for both signal and noise. There are basically four parameters:

π_A - relative abundance of class A

τ - separation of the (mean) signals for the two classes

b_N - a spatial autocorrelation parameter for the noise

ν - a boundary intensity parameter for the signal,

In terms of the preceding specification, the probability that an $N\times N$ neighborhood is class homogeneous, given that its central point x falls in class A, is

$$G_N(1) - G_N(1^-) = 1 - [1-\pi_A]\nu N .$$

A similar calculation for $N=1$ gives the probability p_m that a pixel is not homogeneous, averaged over all classes, viz.,

$$p_m = \nu \cdot \sum_{\alpha} \pi_\alpha(1-\pi_\alpha) \ .$$

If, by definition, we say that inhomogeneous pixels are always misclassified then p_m is a lower bound for the overall classification error rate. The expression for p_m is written to show its generalization to problems with more than two classes.

6. BIBLIOGRAPHY

Kendall, M. G. and Moran, P. A. P. (1963). Geometric Probability, Griffin.

Rosenfeld, A. and Kak, A. C. (1982). Digital Picture Processing, Academic Press.

Switzer, P. (1980). "Extensions of linear discriminant analysis for statistical classification of remotely sensed satellite imagery," Mathematical Geology, 12, 367-376.

Switzer, P., Kowalik, W. S., and Lyon, R. J. P. (1982). "A prior probability method for smoothing discriminant analysis classification maps," Mathematical Geology, 14, 433-444.