ITERATIVE BAYESIAN CONTEXTUAL CLASSIFICATION OF REMOTELY SENSED DATA

Ruben Klein
IMPA, Rio de Janeiro
and
University of California
Riverside

S. James Press
University of California
Riverside

ABSTRACT

In this paper we present an iterative procedure for classifying pixels on the basis of multivariate observations (reflectances) taken by satellite. The method requires the availability of an initial classification. The procedure is multistage, involving successive reclassification of the pixels until the procedure stabilizes. The updating procedure is predictive Bayesian, and contextual. We illustrate the procedure with an example based upon simulated data, and we study the distribution of the simulations.

Key words: Iterative, Bayes, classification, contextual classification, remotely sensed data.

1. Introduction

In this paper we present an iterative procedure for classifying pixels on the basis of p-variate observations (reflectances) taken by satellite. The method requires the availability of an initial classification of the pixels in a scene of interest. The procedure is multistage, involving successive reclassifications of the pixels in the scene of interest, until the procedure stabilizes (the newly reclassified scene, or map, hardly differs from the last stage of classification). The updating procedure is predictive Bayesian, and contextual, in that it depends upon observations from the immediate neighbors of a pixel to be classified, as well as upon the observation vector from that pixel. The method we propose is also "predictive", in that the classification procedure is conditioned on the observed data (the observation vectors, as well as any training data available). When population parameters are known, one way to obtain an initial classification is to use the classical (non-contextual) approach (using a linear discriminant rule, in the case of normal data and equal covariance matrices; or a quadratic discriminant rule, in the case of normal data and unequal covariance matrices). (In this case, frequentist and Bayesian methods coincide). This approach generally yields poor results with spatially correlated data. A superior way would be to use a predictive Bayesian contextual procedure (see Klein & Press, 1989, 1990, 1990b).

Recent work most closely related to ours includes that of Besag, 1986; Fu and Yu, 1980; Haslett, 1985; Hjort and Mohn, 1984; Kittler & Foglein, 1984; Kittler and Pairman, 1985; Mardia, 1984; Owen, 1984; Saebo et al., 1985; Swain, Vardeman, and Tilton, 1981; Switzer, 1980; Tilton, Vardeman, and Swain, 1982; Welch and Salter, 1971; and Yu and Fu 1983.

We describe the iterative classification procedure in Section 2. We show a simulated example in Section 3, and in Section 4 we present some general conclusions.

2. Iterative classification

$Z(s) : p \times 1$ denotes a realization of the covariance stationary, spatial stochastic process $\{Z(s) : s \in R^2\}$, where s is a point on the ground measured in a two-dimensional coordinate system. Denote the observation to be classified as $z_0 \equiv z \equiv (s_0)$. z_0 is to be classified into one of the populations $\pi_i \equiv N(\theta_i, \Sigma_i)$, $i = 1, \ldots, K$, where $\Sigma_i > 0$, which means that Σ is a square, positive definite, symmetric matrix. We assume, along with Elphinstone et al., 1985, that the normality assumption is quite reasonable for LANDSAT type data, and that the covariance matrices are generally unequal.

We consider the immediate neighbor configuration shown in Figure 2.1 (this is the case used in our example in Section 3). In Figure 2.1, z_j denotes an observation vector from the jth immediate neighbor, $j = 1, \ldots, r$, for an $r$neighbor configuration, where we have taken $r = 8$. The population memberships of the r neighbors, relative to that of z_0 is denoted by the conditional configuration probability

$$P\{z_1 \in \pi_{\rho_1}, \ldots, z_r \in \pi_{\rho_r} | z_0 \in \pi_k\}.$$

Let τ_k denote the posterior predictive classification probability for the event that z_0 belongs to π_k. That is,

$$\tau_k \equiv P\{z_0 \in \pi_k | z_0, N, D\},$$

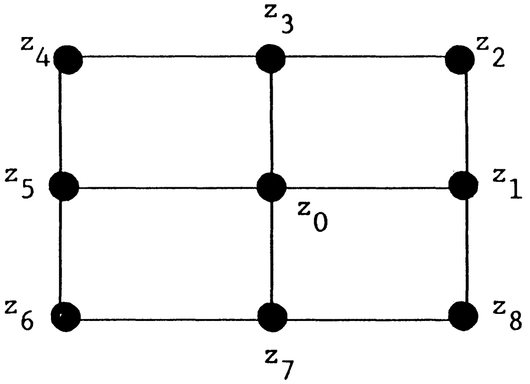

Figure 2.1. 8-neighbor configuration.

where N denotes the observations from the immediate neighbors of z_0, and D denotes the training data observations. By Bayes theorem, using $\propto$ to denote proportionality,

$$r_k \propto p_k(s_0) g\left(z_0, N, D | z_0 \in \pi_k\right),$$

where $p_k(s_0)$ denotes the prior classification probability (which may depend upon the location s_0),

$$p_k(s_k) \equiv P\left\{z_0 \equiv Z(s_0) \in \pi_k\right\},$$

and $g(\cdot)$ denotes the joint probability density for (z_0, N, D), given that $z_0 \in \pi_k$. Conditioning on D, and on the classifications of the neighbors, gives

$$r_k \propto p_k(s_0) \sum_{\rho_1=1}^{K} \cdots \sum_{\rho_r=1}^{K} f\left(z_0, N | D, z_0 \in \pi_k, z_1 \in \pi_{\rho_1}, \ldots, z_r \in \pi_{\rho_r}\right)$$
$$\cdot P\left\{z_1 \in \pi_{\rho_1}, \ldots, z_r \in \pi_{\rho_r} | z_0 \in \pi_k\right\}, \qquad (2.1)$$

where $f(\cdot)$ denotes the joint probability density for z_0 and its neighbors, conditional upon the training data, and upon the classifications of (z_0, N); the second term in the sum is the conditional configuration probability discussed above.

We now assume that somehow, the map has been initially classified, and we want to proceed to stage two for a reclassification. We will use egn. (2.1) to reclassify each pixel in the map. A problem is that we don't actually know the conditional configuration probability; so we approximate it using the initial classification. Call the approximation $\widehat{P}$. P uses the proportions of pixels that estimate P in the initial classification. The larger the scene, the more accurate will be the approximation. While we don't require a Markov random field assumption, we do require a spatial ergodicity assumption to hold, so that spatial averages tend to population averages.

If we fix the classifications of the immediate neighbors at their population memberships established by the initial classification, the summations in eqn.(2.1) disappear, and eqn.(2.1) is approximated by

$$r_k \propto p_k(s_0) f\left(z_0, N | D, z_0 \in \pi_k, z_1 \in \pi_{\rho_1}, \ldots, z_r \in \pi_{\rho_r}\right)$$
$$\cdot \widehat{P}\left\{z_1 \in \pi_{\rho_1}, \ldots, z_r \in \pi_{\rho_r} | z_0 \in \pi_k\right\}. \qquad (2.2)$$

Next note that $f(\cdot)$ may be rewritten as

$$\begin{aligned}&f\left(z_0, N|D, z_0 \in \pi_k, z_1, \in \pi_{\rho_1}, \dots, z_r \in \pi_{\rho_r}\right)\\ =&f_1\left(z_0|N, D, z_0 \in \pi_k, z_1 \in \pi_{\rho_1}, \dots, z_r \in \pi_{\rho_r}\right)\\ \cdot&f_2\left(N|D, z_0 \in \pi_k, z_1 \in \pi_{\rho_1}, \dots, z_r \in \pi_{\rho_r}\right).\end{aligned}$$

But the conditional density of N, $f_2(\cdot)$, depends only upon the classification of the neighbors, and not upon the event $\{z_0 \in \pi_k\}$, so $f_2(\cdot)$ does not depend upon k; so it may be absorbed into the proportionality constant. Eqn.(2.2) then becomes:

$$\begin{aligned}\pi_k \propto\, & p_k(s_0) f_1(z_0|N, D, z_0 \in \pi_k, z_1 \in \pi_{\rho_1}, \dots, z_r \in \pi_{\rho_r})\\ &\cdot P\{z_1 \in \pi_{\rho_1}, \dots, z_r \in \pi_{\rho_r}|z_0 \in \pi_k\}. \qquad (2.3)\end{aligned}$$

What should we use for $p_k(s_0)$ at stage 2? Bayes theorem asserts that the best information we have about z_0 at stage 2 is what we learned from stage 1. Accordingly, at stage 2 we replace $p_k(s_0)$ by τ_k at stage 1, the stage 1 posterior probability for the population membership of z_0. Thus, if $\tau_k(n)$ denotes τ_k at stage n, as an updating equation we use

$$\begin{aligned}\tau_k(n) = \tau_k(n-1) & f_1(z_0|N, D, z_0 \in \pi_k, z_1 \in \pi_{\rho_1}, \dots, z_r \in \pi_{\rho_r})\\ &\cdot \widehat{P}\{z_1 \in \pi_{\rho_1}, \dots, z_r \in \pi_{\rho_r}|z_0 \in \pi_k\}. \qquad (2.4)\end{aligned}$$

Note that $f_1(\cdot)$ and $\widehat{P}(\cdot)$ both may change at each stage of iteration. At any stage we use eqn.(2.4) to reclassify every pixel in the map on the basis of the classification on the previous stage. In the example in Section 3 we use an 8neighbor configuration for reclassification. The procedure is iterated until the new map classification has not changed much from the last classification. In all examples we have tried, the iterative procedure has stabilized at a final map classification in less than 15 iterations (for one-dimensional data assumed to be independent; a longer period of iteration may be required for correlated, high dimensional data). Various loss function criteria may be used for actual classification of a pixel, but a simple criterion to use is: classify a pixel into that π_k for which τ_k is a maximum.

3. Example

In this section we present an example of the interactive classification procedure described in Section 2. The example is based upon simulated data, so we know what the correct map really is. The simulations we carried out differ from most other simulated examples in this area in that instead of running our problem for one or several trials, we have run it for 100 trials, and we have studied the distribution of the simulated results. As a consequence, we can study the variability of the classification procedure, as well as its mean or median percentage of correct classification, for each of the populations present in the scene. We assume observations are one-dimensional, and are independent, given the map. Our criterion for a good classification procedure is: percentage of correct classification (PCC). We also evaluate the PCC for each population separately.

The example involves the 88 x 100 true map shown in Figure 3.1. It is taken from Besag, 1986. The map involves only two populations, but the boundaries between the populations are quite complex. There are 6,398 pixels in $\pi_1 \equiv N(0,1)$ and 2,402 pixels in $\pi_2 \equiv N(\delta, 1)$. We have taken training data from all of the pixels in columns 15, 40, 65, 90. We have used 8 neighbors for updating, as well as 8 neighbors to arrive at an initial predictive Bayesian contextual classification (see Klein and Press, 1989).

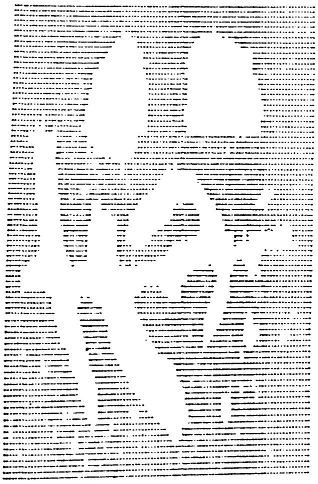

Figure 3.1. True Map.

Figure 3.2 shows a comparison of the iterative solution (vertical bars), and the predictive Bayesian 8-neighbor contextual classification solution (blank bars). (Note that the horizontal bar solution in Figure 3.2, arrived at by simulation is approximately the theoretical probability of correct classification for the case of known parameters.) It is seen that the mean PCC always improves with the iterative Bayesian classification solution.

Figure 3.3 shows the combined effects of separation between the populations and quality of initial solution.

Figure 3.4 shows Box-type plots of the simulation distributions.

Figures 3.5a, b, c show, for $\delta = 1$, map classifications corresponding to the means of the simulation distributions. Comparison of Figure 3.5c with the true map in Figure 3.1 shows that the iterative Bayesian classification procedure did an excellent job of classifying the pixels in this complex scene, and of reconstructing the original image (about 92% of the pixels were correctly classified).

Our final graph, Figure 3.6, shows Box-type plots for the case of $\pi_1 \equiv N(0,1)$, and $\pi_2 \equiv N(1,1)$. The results are seen to be extremely stable with little variability about a PCC in the vicinity of 90%.

4. Conclusions

(1) The Iterative Bayesian solution is better when it starts with an initial contextual solution with 8 neighbors than when it starts with a classical solution (with no neighbors). The difference is about 5% for $\delta = 0.5$, in the example. The conclusion therefore is that, how good the initial classification is matters greatly.

(2) The spread of the distribution of percentage of correct classifications decreases with increasing size of the map, and with increasing separation of the populations.

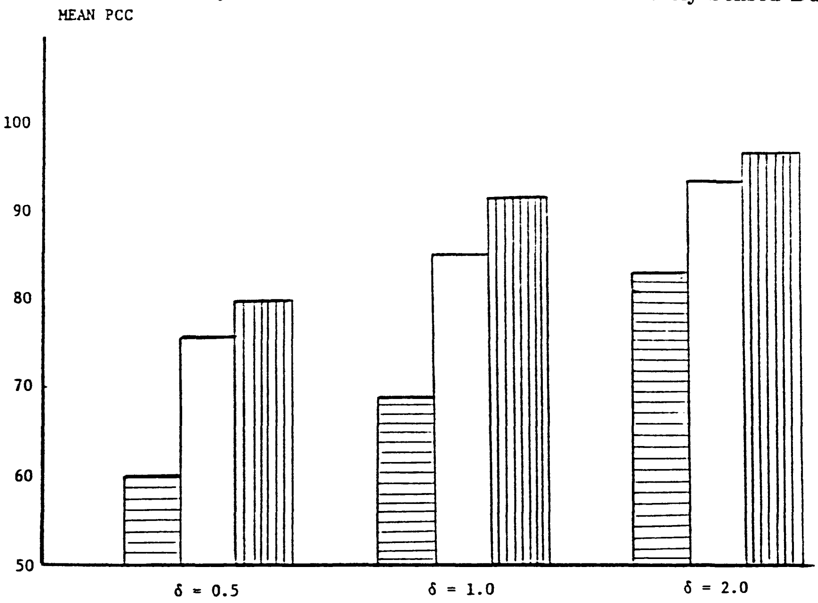

Figure 3.2. Comparison of percentage of correct classifications for iterative solution with classical and 8-neighbor contextual solutions. δ = distance between means. Bars with horizontal ruling: no neighbors (classical solution – no iteration). Blank bars: 8 neighbors (no iteration). Bars with vertical ruling: iterative solution with 8 neighbors, initial solution 8 neighbors.

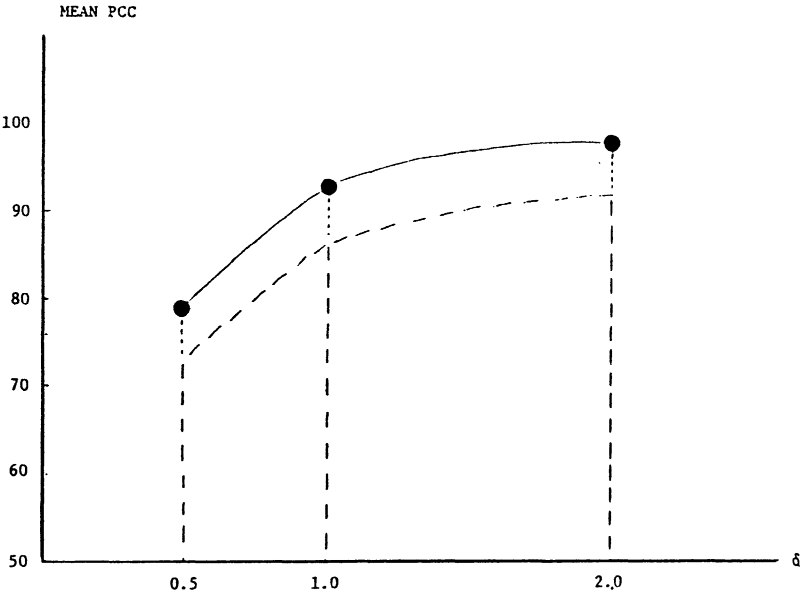

Figure 3.3. Dependence of iterative solution on quality of initial classification. δ = distance between means. PCC = percentage of correct classification. Continuous line: initial solution (8 neighbors). Dashed line: initial solution (no neighbors).

(3) Improvement over the initial contextual solution with 8 neighbors increases with decreasing separation of the populations. Improvement is about 6% for $\delta = 0.5$.

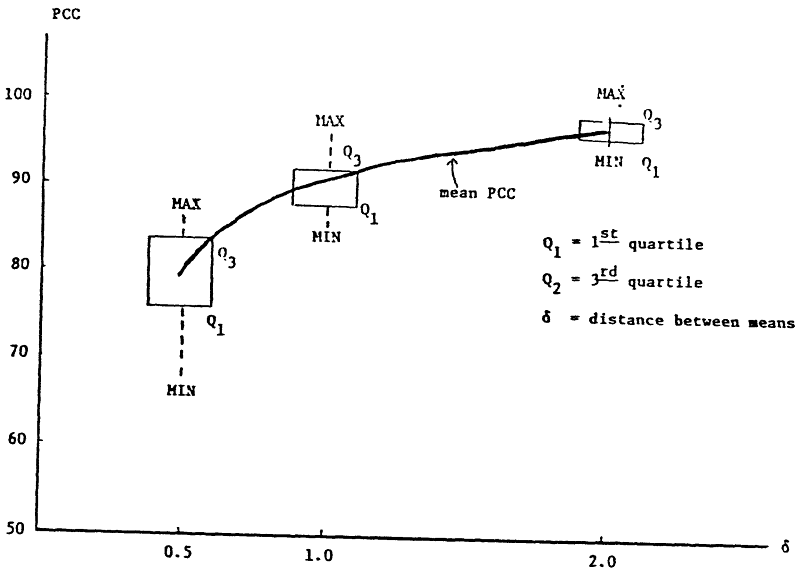

Figure 3.4. Box-type plots for simulation distribution of total percentage of correct classifications.

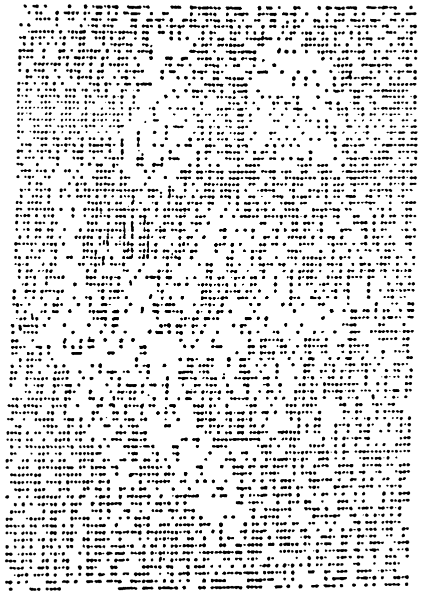

Figure 3.5a. Classical solution – (no neighbors) (PCC, PCC(1), PCC(2)) = (66.5, 63.6, 74.2).

(4) The solution improves over the initial classification and gives a "smoother" map, in the sense of fewer boundary changes that are not required.

(5) The sampling distribution of PCC is approximately the same, regardless of whether we are interested in π_1, π_2, or the overall classification.

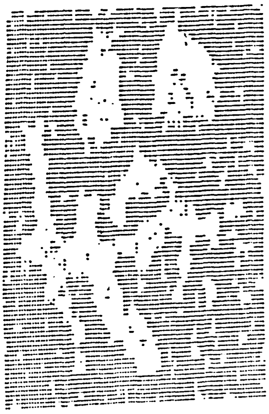

Figure 3.5b. Predictive Bayesian contextual solution – (8 neighbors) (PCC, PCC(1), PCC(2)) = (89.0, 92.6, 79.6).

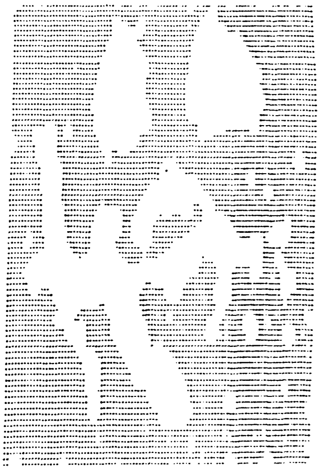

Figure 3.5c. Iterative solution – (with 8 neighbors for updating, and with initial solution with 8 neighbors) (PCC, PCC(1), PCC(2)) = (92.8, 93.2, 91.7).

Acknowledgements

The first author would like to gratefully acknowledge the support of CNPq (Brazil) proc: 20.0213/87.3.

We would like to thank Telma S. Klein for helping us to design the simulation programs used to obtain the results in Section 4. We want to acknowledge the support of the UCR block grant for using the San Diego Supercomputer Center (SDSC).

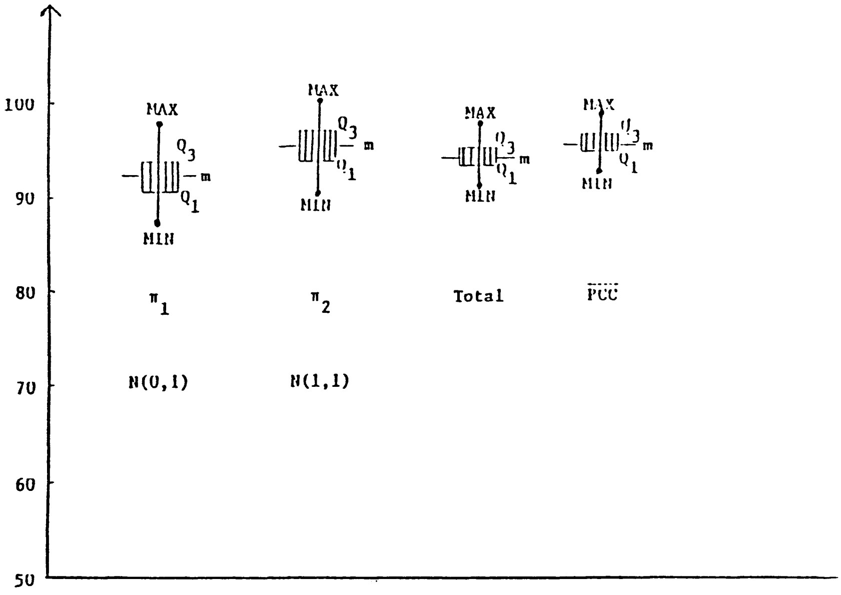

Figure 3.6. Box-type plots for simulation distribution of percentage of correct classification supposing equal variances. Q_1 = 1st quartile, Q_3 = 3rd quartile. Classical PCC = (69.1, 69.1). $\overline{\text{PCC}} = \frac{1}{2}[\text{PCC}(\pi_1) + \text{PCC}(\pi_2)]$

References

Besag, J. (1986). On the Statistical Analysis of Dirty Pictures. *JRSS* B **48** (3), 259–302.

Elphinstone, C. D., Lanergan, A. T., Fatti, L. F., & Hawkins, D. M. (1985). *An Empirical investigation into the Application of Some Statistical Techniques to the Classification of Remotely Sensed Data.* CSIR Special Report SWISK 40, Pretoria, South Africa.

Fu, K. S., & Yu, T. S. (1980). *Statistical Pattern Classification Using Contextual Information.* Research Studies Press, Chichester, UK.

Geisser, S. (1964). Posterior Odds for Multivariate Normal Classifications. *JRSS* B **26**, 69–76.

Geisser, S. (1966). *Predictive Discrimination.* Multivariate Analysis, ed. P. R. Krishnaiah. New York: Academic Press, 149–163.

Haslett, J. (1985). Maximum Likelihood Discriminant Analysis on the Plane Using a Markovian Model of Spatial Context. *Pattern Recognition* (18), Numbers 3 & 4, 287–296.

Hjort, N. L., & Mohn, E. (1984). *A Comparison of Some Contextual Methods in Remote Sensing Classification.* Proceedings of the 18th International Symposium on Remote Sensing of Environment, Volume III, 1693–1702.

Kittler, J., & Foglein, J. (1984). *Contextual Classification of Multispectral Pixel Data.* Image and Vision Computing, Volume 2, 13–29.

Kittler, J., & Pairman, D. (1985). Contextual Pattern Recognition Applied to Cloud Detection and Identification. *IEEE Transactions on Geoscience and Remote Sensing* GE–23 **6**, 855–863.

Klein, R., & Press, S. J. (1989). Contextual Bayesian Classification of Remotely Sensed Data. *Communications in Statistics-Theory and Methods* **18** (9), 3177–3202.

Klein, R., & Press, S. J. (1990). *Bayesian Contextual Classification with Neighbors Correlated with Training Data*, in Bayesian and Likelihood Methods in Statistics and Econometrics: Essays in Honor of George A. Bamard, S. Geisse, J. Hodges, S. J. Press, A. Zellner (eds.) New York, North Holland Publishing Co., 337–355.

Klein, R., & Press, S. J. (1990b). *Contextual Bayesian Classification of Remotely Sensed Data when Training Data is Part of the Scene.* Revista Brasileira De Probabilidade E Estatistica, in English, Volume 4, No.1.

Mardia, K. V. (1984). Spatial Discrimination and Classification Maps. *Communications in Statistics, Theory and Methods* **13** (18), 2181–2197.

Owen, A. (1984). A Neighborhood-Based Classifier for *LANDSAT* Data. *Canadian Hour. of Statistics* **12**, 191–200.

Saebo, H., Braten, K., Hjort, N. L., Llewellyn, B., & Mohn, E. (1985). *Contextual Classification of Remotely Sensed Data: Statistical Methods and Development of a System*, Norwegian Computing Centre Report 768.

Swain, P. H., Vardeman, S. B., & Tilton, J. C. (1981). Contextual Classification of Multispectral Image Data, *Pattern Recognition* **13** (6), 429–441.

Switzer, P. (1980). Extensions of Linear Discriminant Analysis for Statistical Classification of Remotely Sensed Satellite Imagery. *Mathematical Geology* **12**, (4).

Tilton, J. C., Vardeman, S. B., & Swain, P. H. (1982). Estimation of Context for Statistical Classification of Multispectral Image Data. *IEEE Transactions on Geoscience and Remote Sensing* **GE-20** (4), 445–452.

Welch, J. R., & Salter, K. G. (1971). A Context Algorithm for Pattern Recognition and Image Interpretation. *IEEE Transactions on Syst.* **SMC-1** 24–30.

Yu, T. S., & Fu, K. S. (1983). Recursive Contextual Classification Using a Spatial Stochastic Model. *Pattern Recognition* **16**, 89–108.