AgRISTARS

SR-PC-00443
NAS9-15466

A Joint Program for Agriculture and Resources Inventory Surveys Through Aerospace Remote Sensing

Supporting Research

January 1980

Technical Report

Contextual Classification of Multispectral Image Data

by P. H. Swain, S. B. Vardeman, J. C. Tilton

Laboratory for Applications of Remote Sensing
Purdue University
West Lafayette, Indiana 47907

SR-PO-00443
NAS9-15466
LARS 011080

TECHNICAL REPORT

CONTEXTUAL CLASSIFICATION

OF MULTISPECTRAL IMAGE DATA

By

P. H. Swain, S. B. Vardeman, J. C. Tilton

This report describes activity carried out in support of the Area Estimation Research activities of the Supporting Research Project

Purdue University
Laboratory for Applications of Remote Sensing
West Lafayette, Indiana 47906

January 1980

CONTEXTUAL CLASSIFICATION OF MULTISPECTRAL IMAGE DATA

TABLE OF CONTENTS

This paper has been submitted for publication in Pattern Recognition. It is similar in content to Section 2C2 of "Data Preprocessing and Information Extraction," LARS Contract Report 113079, Vol. III, Part II of the Annual Report on NASA Contract No. NAS9-15466, Laboratory for Applications of Remote Sensing, Purdue University, West Lafayette, IN 47907, November 1979.

LIST OF FIGURES

LARS Technical Report 011080

CONTEXTUAL CLASSIFICATION OF MULTISPECTRAL IMAGE DATA

By

P. H. Swain, S. B. Vardeman, J. C. Tilton

ABSTRACT

Compound decision theory is invoked to develop a model for classifying image data using spatial context. Methods for characterizing contextual information in an image are proposed and tested. Experimental results based on both simulated and real multispectral remote sensing data demonstrate the effectiveness of the contextual classifier. A number of practical problems associated with this approach are discussed and possible solutions are explored.

This research was funded in part by the National Aeronautics and Space Administration, Contract No. NAS9-15446, and the National Science Foundation, Grant MCS78-04366.
Authors Swain and Tilton are with the School of Electrical Engineering and Laboratory for Applications of Remote Sensing, Purdue University; Vardeman is with the Department of Statistics, Purdue University, West Lafayette, IN 47907.

1. INTRODUCTION

Multispectral image data collected by remote sensing devices aboard aircraft and spacecraft are relatively complex data entities. Both the spatial attributes and spectral attributes of these data are known to be information bearing[1], but to reduce the magnitude of the computations involved, most analysis efforts have focused on one or the other. Only within the last few years have serious efforts been made to utilize them jointly. For example, one approach uses the spectral homogeneity of "objects," such as agricultural fields, to segment the scene and then uses sample classification to assign each object as a whole, rather than its individual pixels (picture elements), to an appropriate ground cover class[2]. Another approach involves extraction of features based on gray-tone spatial-dependency matrices from which texture-like characteristics are developed[3].

In this paper we describe a more general way to exploit the spatial/spectral context of a pixel to achieve accurate classification. Just as in written English one can expect to find certain letters occurring regularly in particular arrangements with other letters (qu, ee, est, tion), so certain classes of ground cover are likely to occur in the "context" of others. The former phenomenon has been used to improve character recognition accuracy in text reading machines. We shall demonstrate that the latter can be used to improve accuracy in classifying remote sensing data. Intuitively this should not be surprising since one can easily think of ground cover classes more likely to occur in some contexts than in others. One does not expect to find wheat growing in the midst of a housing subdivision, for example. A close-grown lush vegetative cover in such a location is more likely the turf of a lawn.

2. THE MODEL

Consistent with the general characteristics of imaging systems for remote sensing, we assume a two-dimensional array of $N = N_1 \times N_2$ pixels of fixed but unknown classification, as shown in Figure 1.

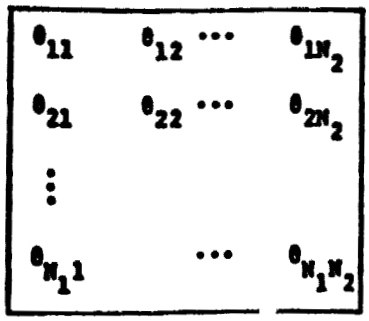


Figure 1. A two-dimensional array of $N = N_1 \times N_2$ pixels.

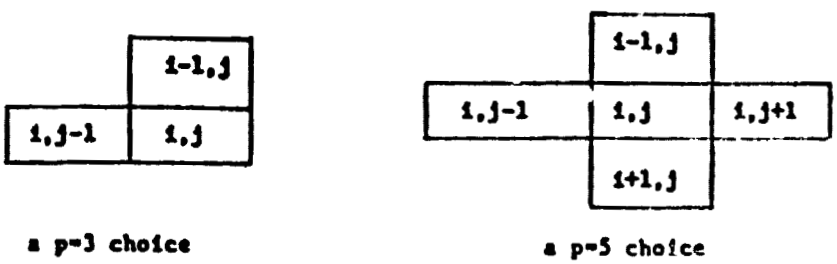


Figure 2. Examples of p-context arrays.

Associated with the pixel having image coordinates (i,j) is its true state or true classification $\theta_{ij} \varepsilon \Omega = \{\omega_1, \omega_2, \ldots, \omega_m\}$, and a random measurement vector (observation) $X_{ij} \varepsilon R^n$ having class-conditional density $p(X_{ij}|\theta_{ij})$. We note that $\{p(X|\omega_i), \quad i=1,2,\ldots,m\}$ is the set of class-conditional probability density functions associating the multispectral measurement vector X with the classes.

Let $\underline{X}$ denote a vector whose components are the ordered pixel measurement vectors:

$$\underline{X} = [X_{ij} | i=1,2,\ldots,N_1; \; j = 1,2,\ldots,N_2]^T .$$

Similarly, let $\underline{\theta}$ be the vector of states:

$$\underline{\theta} = [\theta_{ij} | i=1,2,\ldots,N_1; \; j=1,2,\ldots,N_2]^T .$$

The individual measurement vectors are assumed to be class-conditionally independent; that is, their joint density can be written as:

$$p(\underline{X}|\underline{\theta}) = \prod_{i,j} p(X_{ij}|\theta_{ij}). \qquad (1)$$

Evidence that this is a reasonable assumption may be found in reference[4].

Let the action (classification) taken with respect to pixel (i,j) be denoted by $a_{ij} \varepsilon \Omega$. The loss suffered by taking action a_{ij} when the true class is θ_{ij} is denoted by $L(\theta_{ij}, a_{ij})$ for some fixed non-negative function $L(.,.)$. Then the average loss suffered over the N classifications in

the array is

$$L = \frac{1}{N}\sum_{i,j} L(\theta_{ij}, a_{ij}).$$

If we make the action a_{ij} a function of the observations, then for a given array $\underline{\theta}$ the expected average loss (or risk) is

$$R_{\underline{\theta}} = E\left[\frac{1}{N}\sum_{i,j} L(\theta_{ij}, a_{ij}(\underline{X}))\right] \qquad (2)$$

where the expectation is with respect to the distribution of the vector of observations.

Our objective may be stated as follows: We want to determine the dependence of the decision function $a_{ij}(\cdot)$ on $\underline{X}$ in such a way that for any given array $\underline{\theta}$, the risk, equation (2), will be minimum. One way to approach the problem of making $R_{\underline{\theta}}$ small is to view $\underline{\theta}$ as a realization of a random process in two dimensions and to derive a decision rule which is Bayes versus this "prior distribution" for $\underline{\theta}$ (probably under some simplifying assumptions concerning the nature of this process). This is the approach of Welch and Salter[5] and Yu[6], who make assumptions on the random process sufficient to guarantee that the Bayes decision concerning pixel (i,j) depends on $\underline{X}$ only through X_{ij} and the four nearest neighbors of the pixel.

We will adopt an approach to controlling $R_{\underline{\theta}}$ through $a_{ij}(\cdot)$ that is more closely related to the large body of statistical literature traceable to Robbins[7], and

known as compound decision theory. See, for example, the works and references of VanRyzin[8,9], Cover and Shenhar[10], and Vardeman[11,12]. Rather than looking for a distribution for $\underline{\theta}$ whose associated Bayes rule is both simple and has small $R_{\underline{\theta}}$ for most $\underline{\theta}$, we use the following argument. First, specify some arrangement of p pixel locations including a pixel to be classified. Call this arrangement the p-context array, several choices of which are shown in Figure 2.

Let $\underline{\theta}^p \epsilon \Omega^p$ and $\underline{X}^p \epsilon (R^n)^p$ stand respectively for p-vectors of classes and n-dimensional measurements; each component of $\underline{\theta}^p$ is a variable which can take on values in Ω; each component of $\underline{X}^p$ is a random n-dimensional vector which can take on values in the observation space. Correspondence of the components of $\underline{\theta}^p$ and $\underline{X}^p$ to the positions in the p-context array is fixed but arbitrary except that the pixel to be classified in the array will always correspond to the pth components. The notation $\underline{\theta}_{ij}$ and $\underline{X}_{ij}$ will refer to the particular instance of $\underline{\theta}^p$ and $\underline{X}^p$ associated with pixel (i,j).

Now consider finding an optimal decision rule of the form

$$a_{ij}(\underline{X}) = d(\underline{X}_{ij}) \tag{3}$$

for a fixed function $d(\cdot)$ mapping p-vectors of observations to actions. The risk associated with any rule of this form is, from equation (2),

$$R_{\underline{\theta}} = E\left[\frac{1}{N}\sum_{i,j} L(\theta_{ij}, d(\underline{X}_{ij}))\right]$$

$$= \frac{1}{N}\sum_{i,j} E\left[L(\theta_{ij}, d(\underline{X}_{ij}))\right]$$

$$= \sum_{\underline{\theta}^p \epsilon \Omega^p} G(\underline{\theta}^p) E[L(\theta_p, d(\underline{X}^p))] \qquad (4)$$

where $G(\underline{\theta}^p)$, the context distribution, is the relative frequency with which $\underline{\theta}^p$ occurs in the array $\underline{\theta}$ and θ_p is the pth component of $\underline{\theta}^p$. Notice that $R_{\underline{\theta}}$ depends on $\underline{\theta}$ only through $G(\underline{\theta}^p)$. Writing equation (4) in more detail and invoking the class-conditional independence assumption, equation (1), we have

$$R_{\underline{\theta}} = \sum_{\underline{\theta}^p \epsilon \Omega^p} G(\underline{\theta}^p) \int L(\theta_p, d(\underline{X}^p)) \prod_{i=1}^{p} p(X_i | \theta_i) d\underline{X}^p$$

$$= \int \sum_{\underline{\theta}^p \epsilon \Omega^p} G(\underline{\theta}^p) L(\theta_p, d(\underline{X}^p)) \prod_{i=1}^{p} p(X_i | \theta_i) d\underline{X}^p \qquad (5)$$

where the product is over the components X_i of $\underline{X}^p$. For any array $\underline{\theta}$, a decision rule $d(\underline{X}^p)$ minimizing $R_{\underline{\theta}}$ can be obtained by minimizing the integrand in equation (5) for each $\underline{X}^p$; thus for a specific $\underline{X}_{ij}$ (an instance of $\underline{X}^p$), an optimal action is:

$d(\underline{X}_{ij})$ = the action (classification) a which minimizes

$$\sum_{\underline{\theta}^p \epsilon \Omega^p} G(\underline{\theta}^p) L(\theta_p, a) \prod_{i=1}^{p} p(X_i | \theta_i).$$

This can be written in a slightly different form which makes more apparent the specific contribution due to context (the term in brackets below):

$d(\underline{X}_{ij})$ = the action a which minimizes

$$\sum_{\theta' \in \Omega} \left[\sum_{\substack{\underline{\theta}^p \in \Omega^p, \\ \theta_p = \theta'}} G(\underline{\theta}^p) \prod_{i=1}^{p-1} p(X_i|\theta_i) \right] L(\theta', a) p(X_p|\theta'). \quad (7)$$

In practice, a "0-1 loss function" is usually assumed, i.e.,

$$L(\theta, a) = \begin{cases} 0, & \text{if } \theta = a \\ 1, & \text{if } \theta \neq a \end{cases}$$

Then (7) simplifies and the decision rule becomes:

$d(\underline{X}_{ij})$ = the action a which <u>maximizes</u>

$$\left[\sum_{\substack{\underline{\theta}^p \in \Omega^p, \\ \theta_p = a}} G(\underline{\theta}^p) \prod_{i=1}^{p-1} p(X_i|\theta_i) \right] p(X_p|a) \quad (8)$$

Thus (8) defines a set of discriminant functions for the classification problem.

The optimal choice of $d(\cdot)$ cannot actually be determined because it depends on $G(\underline{\theta}^p)$ which is unknown.

We can, however, expect that, at least for large $N = N_1 \times N_2$, a decision rule in which $G(\underline{\theta}^p)$ is replaced by an estimate $\hat{G}(\underline{\theta}^p)$ based on the $\underline{X}_{ij}$ will have risk $\hat{R}_{\underline{\theta}}$ approximating that of the optimal rule. (We call this the "bootstrap effect.") That this is the case when $p = 1$ (approximating an optimal pointwise classifier with estimated a priori probabilities) and suitable forms of estimation are used is a consequence of the work of VanRyzin[9].

The notion of attempting to approximate the risk of the best rule of the form equation (3) for $p > 1$, given its first general treatment in Gilliland and Hannan[13], has not been as thoroughly studied as the $p = 1$ version. But related work for $p > 1$ in sequence versions of compound decision theory[14] suggests the validity of the generalization. Further, Vardeman[12] points out that if one is willing to separate the N locations into several groups $G_1, G_2, \ldots, G_\ell$ within each of which the $\underline{X}_{ij}$ are independent, the results for $p = 1$ by VanRyzin guarantee that, for $p > 1$, replacing the $G(\underline{\theta}^p)$ by estimates of the frequencies of $\underline{\theta}^p$ group-by-group produces a decision procedure having the risk of the optimal rule as an approximate upper bound on its risk. An illustration of this separation idea is shown in Figure 3.

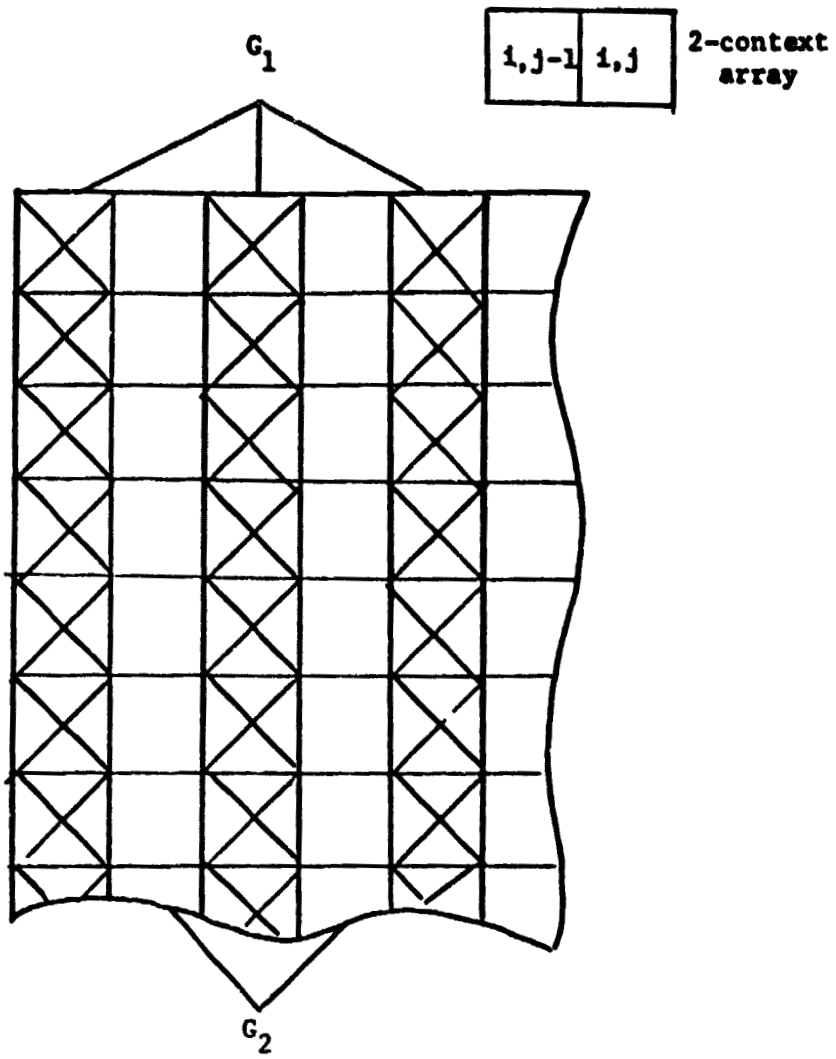


Figure 3. A 2-context array with separable pixel groups.

In the interest of a practical solution to the problem of incorporating context into the classification procedure, estimates of $G(\underline{\theta}^p)$ were derived experimentally by simply counting the occurrences of each $\underline{\theta}^p$ obtained in a preliminary classification of the scene without the use of context. Although the use of this rather crude method of estimating $G(\underline{\theta}^p)$ has not been studied in the statistical literature, we will demonstrate in Section 3 its effectiveness for our application.

Before proceeding to a discussion of our experimental results, we make two further observations concerning this approach. First, seeking a criterion for the "context richness" of a scene, we have been able to reach only the following result. Suppose the frequencies $G(\underline{\theta}^p)$ are such that $G(\underline{\theta}^p)$ can be written in factored form, i.e.,

$$G(\underline{\theta}^p) = G_1(\underline{\theta}').G_2(\underline{\theta}")$$

where $\underline{\theta}'$ and $\underline{\theta}"$ are, respectively, $p - \ell$ and ℓ vectors of classes. Then (6) can be written in the form

$$\sum_{\underline{\theta}"} L(\theta_p, a) \prod_{\substack{i= \\ p-\ell+1}}^{p} p(X_i|\theta_i) G_2(\underline{\theta}") \quad . \quad \sum_{\underline{\theta}'} \prod_{i=1}^{p-\ell} p(X_i|\theta_i) G_1(\underline{\theta}').$$

But now the terms included in the second summation are independent of the conditions at the pixel to be classified and are therefore constant relative to the decision to be made. Thus, the decision depends only on ℓ components of the p-context array and is independent of the other p-ℓ locations. If it were possible to determine such factorability of the $G(\underline{\theta}^p)$, one could simplify the context classification computations by reducing the size of the context array.

Second, comparing (7) with the results of Welch and Salter[5] and reinterpreting the $G(\underline{\theta}^p)$ as the marginal of an a priori distribution for $\underline{\theta}$, one may view (7) as a generalization of the Welch and Salter context classification rule. The advantages of the present formulation are that one need make no possibly unrealistic assumptions about the distribution for $\underline{\theta}$ and has complete freedom to choose both p and the form of the p-context array. There are situations (e.g., locating clouds and their associated shadows in a scene) in which context arrays other than those involving neighboring pixels would be useful, a possibility unique to this approach.

3. EXPERIMENTAL RESULTS

Experiments were performed to explore the effectiveness of contextual classification as applied to the analysis of multispectral remote sensing data. First, simulated data were used to determine the degree to which contextual classification might improve the analysis results (as compared to no-context classification), given that the class-conditional densities and the context distribution for the scene were known. The simulated data were used again to investigate candidate methods for estimating the context distribution since, as noted in Section 2, it usually cannot be assumed that the context distribution is known a priori. Finally, contextual classification was applied to real data to determine the extent to which the conclusions drawn from the simulated-data experiments could be extended to the more realistic case.

Simulated Data Experiments

A no-context classification of multispectral remote sensing data was selected which had been judged to be very accurate (produced by careful analysis and refinement of multitemporal data). Such a classification could be expected to embody the contextual content of an actual ground scene. Using the classification map and the

associated statistics of the classes (developed in performing the no-context classification), data vectors were produced by a Gaussian random number generator and composed into a new data set. Thus the new data set had the following characteristics:

(1) Each pixel in the simulated data set represented the same class as in the "template" classification. The template could be considered the "ground truth" for the simulated data set.

(2) All classes in the data set were known and represented.

(3) All classes had multivariate Gaussian distributions with statistics typical of those found in real data.

(4) All pixels were class-conditionally independent of adjacent pixels.

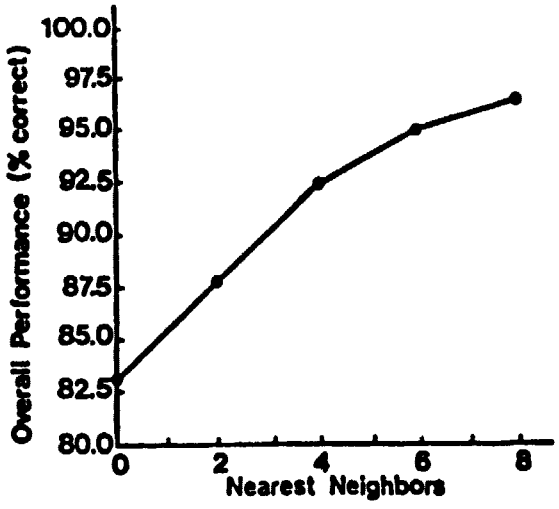
(5) There were no mixture pixels.

Data simulated in this manner are somewhat of an idealization of real remote sensing data, but the spatial organization of the simulated data is consistent with a real world scene and the overall characteristics of the data are consistent with the contextual classifier model. In essence, then, the experimental results based on the simulated data demonstrate the effectiveness of the context classifier, given that the underlying assumptions are satisfied. Further experiments with real data are required to generalize the conclusions.

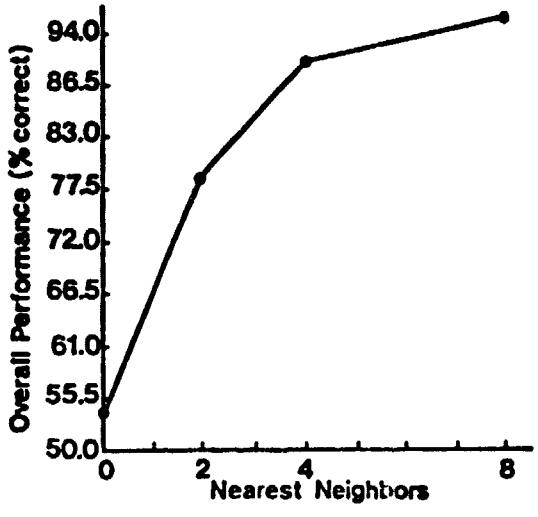
Three data sets were selected representing a variety of ground cover types and textures. Data set 1 was agricultural (Williston, North Dakota), with ground resolution and spectral bands approximating those of the projected Landsat-D Thematic Mapper. Data set 2a was Landsat-1 data from an urban area (Grand Rapids, Michigan). Data set 2b was from the same Landsat frame as 2a, but from a locale having significantly different spatial organization. Each data set was square, 50 pixels on a side.

Figure 4 shows the classification results obtained. The "no-context" classification accuracy is plotted coincident with the vertical axis of each graph. Data set 1 was classified using successively 0, 2, 4, 6 and 8 neighboring pixels; data sets 2a and 2b were classified using 0, 2, 4 and 8 neighboring pixels. The accuracy improvement resulting from the use of contextual information was found to be quite significant.

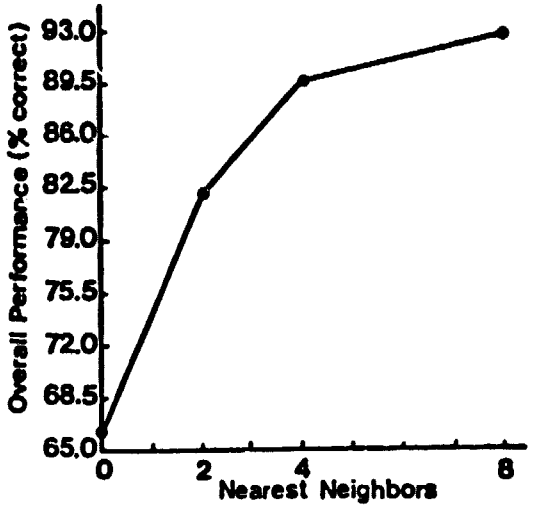
To accomplish the context classification using this approach, it is necessary to have available the class-conditional density functions for the classes to be recognized, $p(X|\omega_i)$, and the context distribution (the frequency distribution associated with the p-vectors,



(a)



(b)



(c)

Figure 4. Contextual classification of simulated data. (a) Data set 1. (b) Data set 2a. (c) Data set 2b.

$G(\underline{\theta}^p)$). In remote sensing applications, the class-conditional density functions are typically learned from training samples. For the experiments described above, the Gaussian class statistics on which the data simulations were based were used for the classification (these were originally the training statistics used to produce the "template" classification). An important question is how in practice to determine the context distribution. In the foregoing experiment, this distribution was simply tabulated from the "template" classification (actually, from an area somewhat larger than classified in this test). But in a real data situation, such a template is not available, else there would be no need to perform any further classification.

One can envision a number of ways in which the context distribution might be estimated for a given remote sensing application. For example, it could be extracted from a classification of data obtained previously from the same area. This would require that the area not have changed much in its class make-up since the earlier data were collected and that the earlier classification was reasonably accurate. Alternatively, the distribution might be obtained from a classification of any similarly constituted area. Still another possibility would be to estimate the context distribution for the data to be classified from a "conventional" classification of the same data determined to have "reasonably good" accuracy.

Conceivably, one might then refine the contextual classification by making another estimate of the context distribution based on the resulting more accurate classification, and even iterate in this way until no further improvements in accuracy were obtained. All of these methods produce an estimate of the context distribution, and a crucial question on which hinges the utility of this contextual classification method is how sensitive the contextual algorithm is likely to be to the "goodness" of the estimate.

The iterative technique starting with a no-context classifiction seemed to be the most practical approach, since no classifications are needed from earlier data or from other areas of similar context. All that is needed is a good initial point-by-point classification of the area in question.

To test the potential of this "bootstrap" technique, it was first tried on the simulated data set 2a. Also, the classifications using the reference template were rerun using an estimate of the context distribution from just the 50-pixel-square area classified, rather than from the larger area (276 x 320) used to obtain the estimate for the results presented in Figure 4. This was done to provide a better comparison to what could be accomplished using the bootstrap technique.

Using this approach, seven iterations (classifications followed by re-estimation of the context distribution) produced an improvement of 36 percent in overall accuracy compared to the point classification using equal a priori

probabilities (from 52 percent to over 86 percent). No significant change was observed in average-by-class accuracy (constant at 68 percent).* This compares with an increase of over 44 percent in overall accuracy (28 percent in average-by-class accuracy) obtained using the context distribution estimated from the template classification. These results are summarized in Figure 5.

As seen in Figure 5, a number of values of p were used in the iteration process. At each iteration, the best classification found by varying p, as judged by trading off overall accuracy against average-by-class accuracy, was used as the template for re-estimating the context distribution for the next iteration.

---

* Classification performance can be tabulated in two ways. Overall accuracy is simply the overall number of correct classifications divided by the total number attempted. Average-by-class accuracy is obtained by first computing the accuracy for each class and taking the arithmetic average of the class accuracies. The latter is significant when the classification results exhibit a tendency to discriminate in favor of or against a subset of the classes.

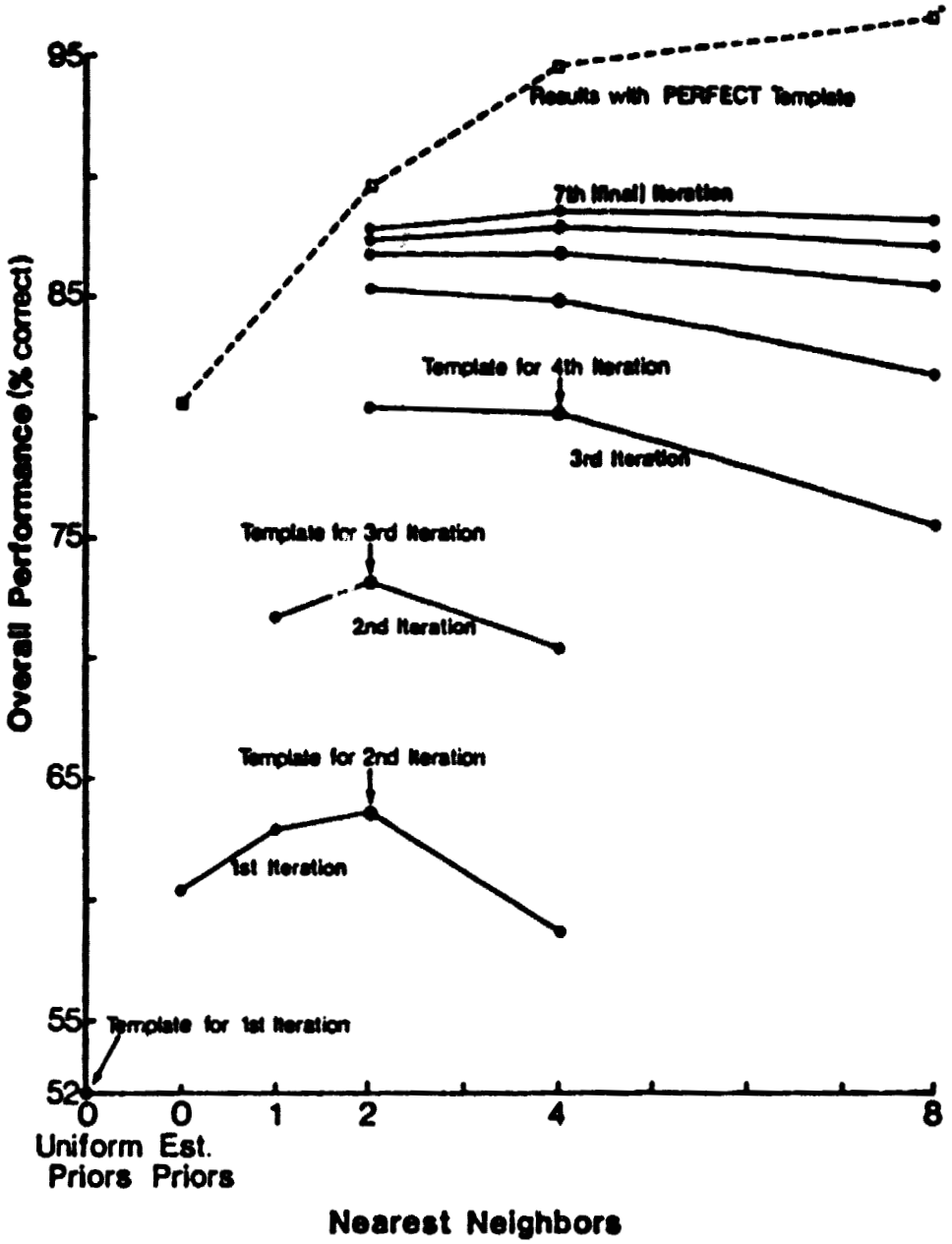


Figure 5. Results of contextual classification using iteratively estimated context distribution (simulated data set 2a).

The best classification on the first iteration was obtained for p = 3 (two nearest neighbors), which was also the case for the second iteration. On the third iteration, the p = 5 (four nearest neighbors) choice was deemed best. Finally, by the seventh iteration, the p = 9 (eight nearest neighbors) choice was considered best. In this case, the overall accuracy was slightly less than for the p = 5 choice (88.2 percent versus 88.6 percent), but the average-by-class accuracy was better by a larger margin (68.1 percent versus 67.4 percent).

This implementation of the bootstrap technique involves a larger number of classifications, usually three or more per iteration. A simpler approach would be to do just one classification per iteration and increase the number of nearest neighbors used for each iteration. As shown in Figure 6, for data set 2a the final result using this method was virtually the same as for the more involved procedure.

It was wondered just how much of the accuracy improvement was due to a better estimate of the point-by-point prior probabilities. After five iterations doing 0-nearest-neighbor classification, the improvement in overall accuracy saturated at 80.3 percent, but the

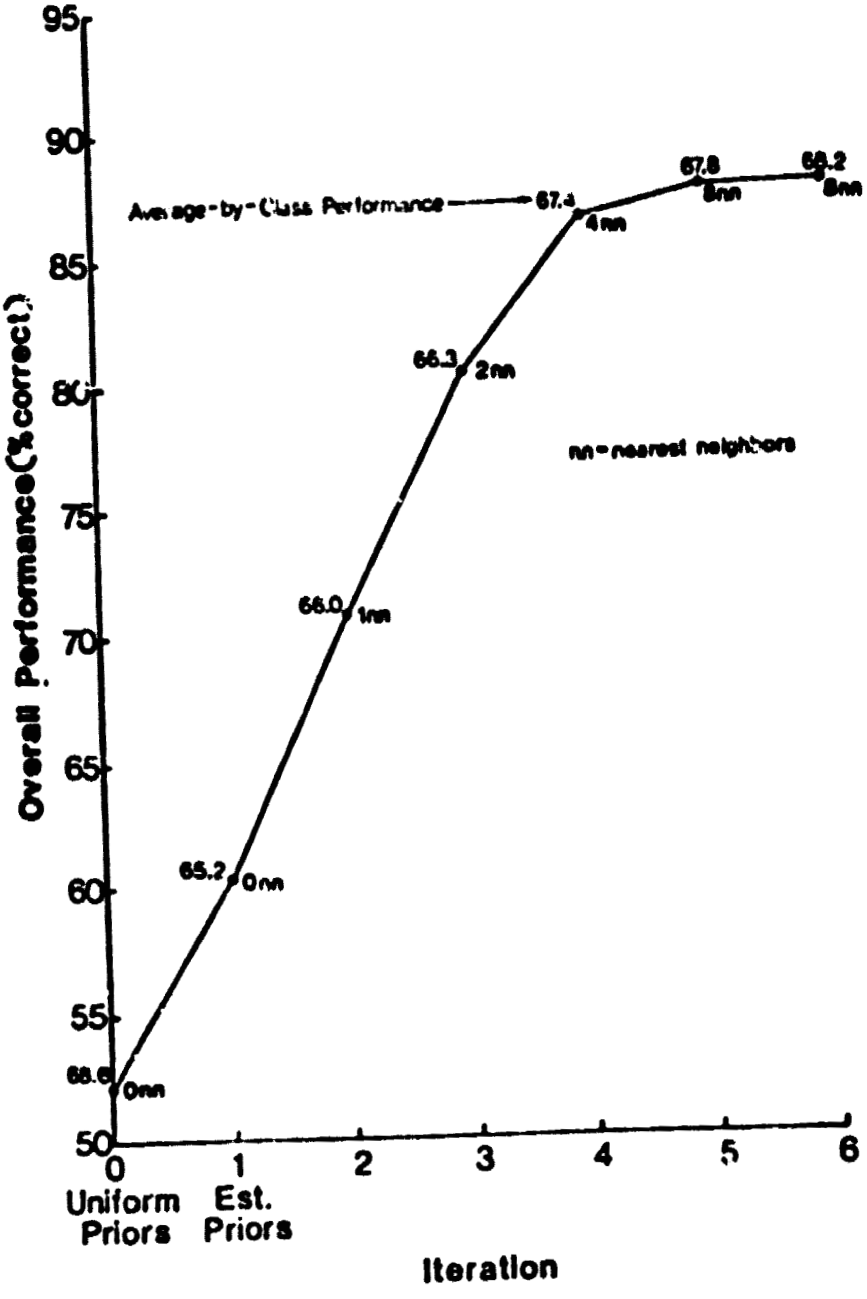


Figure 6. Contextual classification results based on simplified iterative technique (simulated data set 2a).

average performance by class had degraded to 46.9 percent. This compares closely to the 0-nearest-neighbor classification done using the context distribution determined from the reference template, which had an overall accuracy of 80.8 percent and an average performance by class of 48.3 percent. It appears from this result that the context serves to improve the overall performance compared to that of the 0-nearest-neighbor result while resisting degradation in average-by-class accuracy.

Real Data Experiments

Having observed excellent performance of the contextual classifier on simulated data, the next step was to see how well it would perform on real data. A 50-pixel-square segment of Landsat data was chosen which included approximately equal amounts of urban and agricultural area located to the southeast of Bloomington, Indiana. Statistics for the spectral classes were estimated using the 100-pixel-square area centered on the 50-pixel-square segment. A very careful classification using 14 spectral classes was performed to delineate agricultural, urban and forested areas. As there were too few forested pixels to delineate forest test areas reliably, the classification was tested only for accuracy in classifying the agricultural and urban classes. Out of the 2500 pixels in the segment, a total of 867 pixels were manually

interpreted as agricultural and 450 pixels as urban. The identification was made by interpretation of color infrared photography taken by aircraft on the same day as the Landsat pass.

The results from using the full bootstrap technique on this data set were not nearly as favorable as the results obtained from the simulated data. See Figure 7.

The no-context classification using uniform prior probabilities had an overall accuracy of 83.1 percent and an average-by-class accuracy of 82.7 percent. The best classification obtained using this result as a template to estimate the context distribution was a p = 2 (one-nearest-neighbor) classification based on the neighbor to the "north" (85.2 percent overall, 84.7 percent average-by-class). Interestingly, the one-nearest-neighbor result based on the neighbor to the "west" produced a slightly poorer classification (84.2 percent overall, 83.8 percent average by class). No apparent features in the scene would account for the difference (i.e., be seen by eye), raising a new issue yet to be pursued.

The second iteration was performed using the one-nearest-neighbor (north) classification from the first

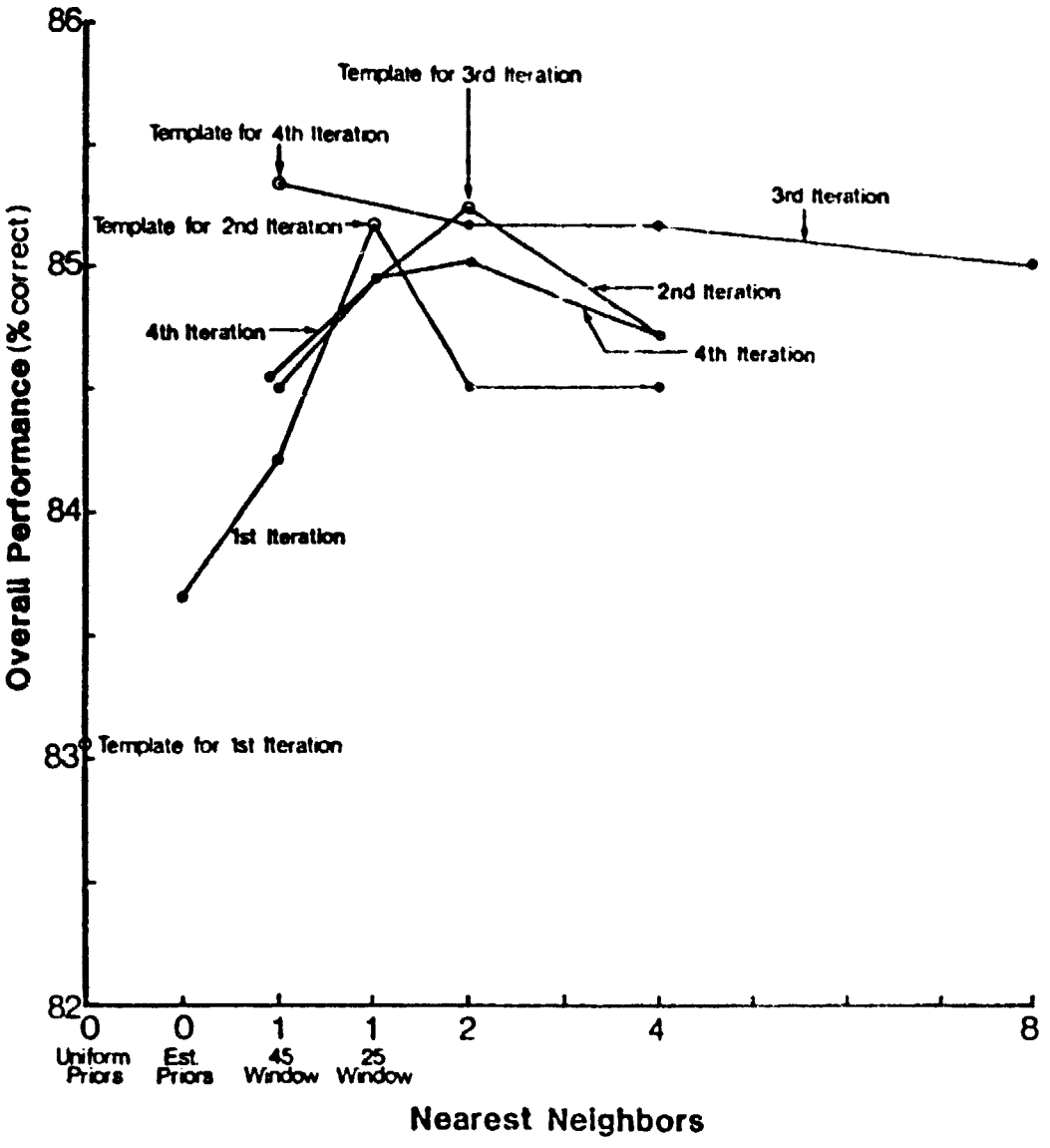


Figure 7. Contextual classification of Bloomington data using the unmodified procedure for estimating the context distribution.

iteration for estimating the context distribution. Here the two-nearest-neighbor (neighbors to the "north" and "west") classification was the best with an overall accuracy of 85.2 percent and average-by-class accuracy of 84.7 percent). The best classification for the third iteration was again the one-nearest-neighbor (north) case with 85.3 percent overall accuracy and 84.8 percent average-by-class accuracy. The fourth iteration produced no improvement. The context classifier thus only yielded just over two percent improvement in both overall accuracy and average-by-class accuracy.

In order to assess the sensitivity of these results to the accuracy of the template used to estimate the context distribution, a manual "cleanup" of the original template was performed, as follows: Change the classification of all incorrectly classified points in the test areas in the original point-by-point uniform priors classification to the closest spectral class in the correct information class as observed by means of a cross-plot of Landsat bands 2 and 3. Where either of two spectral classes might have been the correct class, a coin was tossed to decide the assignment. The context distribution was then estimated from the entire modified classification including both test and non-test areas.

The first iteration using this modified classification as template produced excellent results (Figure 8). The

p = 9 (eight-nearest-neighbor) classification produced an improvement of over 10 percent to 93.8 percent in overall accuracy and over 11 percent to 93.6 percent in average-by-class accuracy (compared to the conventional point classifier with uniform prior probabilities). A second iteration was performed using a context distribution estimated from a similarly modified eight-nearest-neighbors classification from the first iteration. No further improvement in accuracy was observed, suggesting that this iterative process "saturates" very quickly.

Both the "full bootstrap" technique and the manual "cleanup" method were also applied to an agricultural Landsat data set from Kansas. The results were consistent with the results just described for the Bloomington data [16]. The full bootstrap method netted only a two percent improvement in overall accuracy for an eight-nearest-neighbors classification. The manual cleanup of the template classification led to a nine percent improvement (again for eight-nearest-neighbors).

The excellent results produced by using the context distribution estimated from the manually modified point classification suggest the following approach for classification using context:

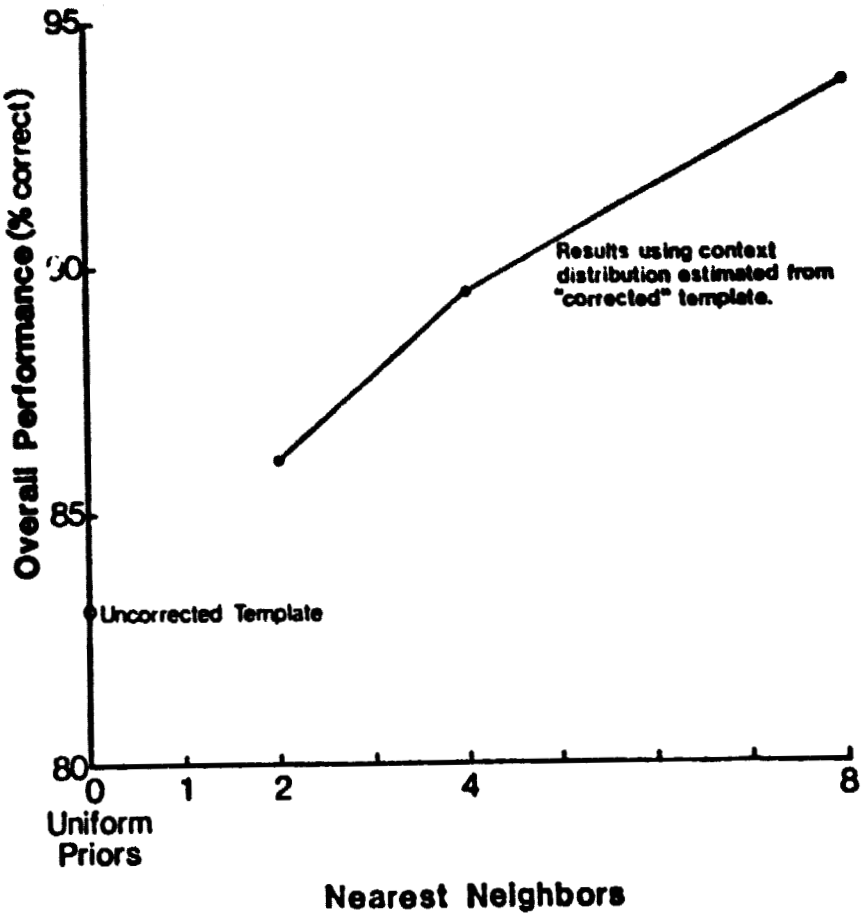


Figure 8. Performance using manual template correction for estimating the context distribution (Bloomington data).

4. SUMMARY AND CONCLUSIONS

The proposed model for a classifier which utilizes contextual information is a generalization of the familiar maximum likelihood classifier. Experimental results based on simulated multivariable data have demonstrated that use of contextual information will significantly improve classification accuracy when the data satisfy the assumptions underlying the classifier model. Results for real data have shown that the obtainable accuracy improvement is dependent, as might be expected, on the accuracy with which the context distribution is known. Although satisfactory results have been achieved, it is clear that further work on ways to improve the context estimation will pay dividends.

The computational demands presented by the contextual classifier are not inconsequential. Fundamentally, the time and space complexity of the method are proportioned to m^p, where m is the number of classes and the context array (including the pixel to be classified) has p cells. Clever implementation schemes are helpful in reducing both the computation time and memory required, but a more practical way to address the problem may be through the use of multiprocessor systems [15]. Measures of "context richness" of a scene would also allow for selective us of the contextual classifier only when significant benefits are likely to be obtained.

REFERENCES

1. P. H. Swain and S. M. Davis, eds., Remote Sensing: The Quantitative Approach, McGraw-Hill International Book Co., New York (1978).

2. R. L. Kettig and D. A. Landgrebe, "Classification of Multispectral Image Data by Extraction and Classification of Homogeneous Objects," IEEE Trans. Geoscience Electronics, Vol. GE-14, pp. 19-26, January (1976).

3. R. M. Haralick. K. Shanmugan, and I. Dinstein, "Textural Features for Image Classification," IEEE Trans. Systems, Man and Cybernetics, Vol. SMC-3, pp. 610-621, November (1973).

4. R. L. Kettig and D. A. Landgrebe, "Computer Classification of Remotely Sensed Multispectral Image Data by Extraction and Classification of Homogeneous Objects," LARS Technical Report 050975, Laboratory for Applications of Remote Sensing, Purdue University, West Lafayette, IN 47907, May (1975).

5. J. R. Welch and K. G. Salter, "A Context Algorithm for Pattern Recognition and Image Interpretation," IEEE Trans. Systems, Man and Cybernetics, Vol. SMC-1, pp. 24-30, January (1971).

6. T. S. Yu and K. S. Fu, "Statistical Pattern Recognition Using Contextual Information," Technical Report TR-EE 78-17, School of Electrical Engineering, Purdue University, West Lafayette, IN 47907, March (1978).

7. H. Robbins, "Asymptotically Subminimax Solutions of Compound Statistical Decision Problems," Proc. Second Berkeley Symp. Mathematical Statistics and Probability, pp. 157-163, University of California Press (1951).

8. J. VanRyzin, "The Compound Decision Problem With m x n Finite Loss Matrix," Annals of Mathematical Statistics, Vol. 37, pp. 412-424 (1966).

9. J. VanRyzin, "The Sequential Compound Decision Problem With m x n Finite Loss Matrix," Annals of Mathematical Statistics, Vol. 37, pp. 954-975 (1966).

10. T. Cover and A. Shenhar, "Compound Bayes Predictors for Sequences With Apparent Markov Structure," IEEE Trans. Systems, Man and Cybernetics, Vol. SMC-7, No. 6, pp. 421-424, (1977).

11. S. Vardeman, "A Note on the Applicability of Sequence Compound Decision Schemes," Scandinavian Journal of Statistics, 6, 2, pp. 86-88 (1979).

12. S. Vardeman, "Solutions to k-Extended Compound Decision Problems, Bootstrap and Bayes," in preparation.

13. D. Gilliland and J. Hannan, "On the Extended Compound Decision Problem," Annals of Mathematical Statistics, Vol. 40, pp. 1536-1541 (1969).

14. J. Ballard, D. Gilliland, and J. Hannan, "$O(N^{-\frac{1}{2}})$ Convergence to k-Extended Bayes Risk in the Sequence Compound Decision Problem with m x n Component," Research Memo RM-333, Statistics and Probability, Michigan State University (1975).

15. P. H. Swain, H. J. Siegel, and B. W. Smith, "Concontextual Classification of Multispectral Remote Sensing Data Using a Multiprocessor System," IEEE Trans. Geoscience Electronics, April (1980).

16. P. H. Swain, P. E. Anuta, D. A. Landgrebe, and H. J. Siegel, "Data Preprocessing and Information Extraction," LARS Contract Report 113079, Vol. III, Part II, Section 2C2, Laboratory for Applications of Remote Sensing, Purdue University, West Lafayette, IN 47907, November 1979.

$$R_{\underline{\theta}} = E\left[\frac{1}{N}\sum_{i,j} L(\theta_{ij}, d(\underline{X}_{ij}))\right]$$

$$= \frac{1}{N}\sum_{i,j} E\left[L(\theta_{ij}, d(\underline{X}_{ij}))\right]$$

$$= \sum_{\underline{\theta}^p \in \Omega^p} G(\underline{\theta}^p) E[L(\theta_p, d(\underline{X}^p))] \qquad (4)$$

where $G(\underline{\theta}^p)$, the <u>context distribution</u>, is the relative frequency with which $\underline{\theta}^p$ occurs in the array $\underline{\theta}$ and θ_p is the pth component of $\underline{\theta}^p$. Notice that $R_{\underline{\theta}}$ depends on $\underline{\theta}$ only through $G(\underline{\theta}^p)$. Writing equation (4) in more detail and invoking the class-conditional independence assumption, equation (1), we have

$$R_{\underline{\theta}} = \sum_{\underline{\theta}^p \in \Omega^p} G(\underline{\theta}^p) \int L(\theta_p, d(\underline{X}^p)) \prod_{i=1}^{p} p(X_i|\theta_i) d\underline{X}^p$$

$$= \int \sum_{\underline{\theta}^p \in \Omega^p} G(\underline{\theta}^p) L(\theta_p, d(\underline{X}^p)) \prod_{i=1}^{p} p(X_i|\theta_i) d\underline{X}^p \qquad (5)$$

where the product is over the components X_i of $\underline{X}^p$. For any array $\underline{\theta}$, a decision rule $d(\underline{X}^p)$ minimizing $R_{\underline{\theta}}$ can be obtained by minimizing the integrand in equation (5) for each $\underline{X}^p$; thus for a specific $\underline{X}_{ij}$ (an instance of $\underline{X}^p$), an optimal action is:

$d(\underline{X}_{ij})$ = the action (classification) a which minimizes

$$\sum_{\underline{\theta}^p \in \Omega^p} G(\underline{\theta}^p) L(\theta_p, a) \prod_{i=1}^{p} p(X_i|\theta_i).$$