E82-10106

NASA-CR-167455

AgRISTARS

SR-P1-04148
NAS9-15466

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

August 1981

Supporting Research

Technical Report

Incorporating Spatial Context Into Statistical Classification of Multidimensional Image Data

James C. Tilton and Philip H. Swain

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

SR-P1-04148
NAS9-15466
LARS 072981

INCORPORATING SPATIAL CONTEXT INTO STATISTICAL CLASSIFICATION OF MULTIDIMENSIONAL IMAGE DATA

James C. Tilton and Philip H. Swain

Purdue University
Laboratory for Applications of Remote Sensing
West Lafayette, Indiana 47906-1399, U.S.A.

August 1981

Star Information Form

1. Report No	2. Government Accession No	3. Recipient's Catalog No
SR-P1-04148		

4. Title and Subtitle	5. Report Date
Incorporating Spatial Context Into Statistical Classification of Multidimensional Image Data	6. Performing Organization Code

7. Author(s)	8. Performing Organization Report No
James C. Tilton and Philip H. Swain	072981

9. Performing Organization Name and Address	10. Work Unit No.
Purdue University Laboratory for Applications of Remote Sensing 1220 Potter Drive West Lafayette, IN 47906-1399	11. Contract or Grant No NAS9-15466

12. Sponsoring Agency Name and Address	13. Type of Report and Period Covered
NASA Johnson Space Center Remote Sensing Research Division Houston, TX 77058	Technical Report 14. Sponsoring Agency Code

15. Supplementary Notes

F.G. Hall, Technical Monitor
M.E. Bauer, Principal Investigator

16. Abstract

Compound decision theory is employed to develop a general statistical model for classifying image data using spatial context. The classification algorithm developed from this model exploits the tendency of certain ground-cover classes to occur more frequently in some spatial contexts than in others. A key input to this contextual classifier is a quantitative characterization of this tendency: the context function. Several methods for estimating the context function are explored, and two complementary methods are recommended. The contextual classifier is shown to produce substantial improvements in classification accuracy compared to the accuracy produced by a non-contextual uniform-priors maximum likelihood classifier when these methods of estimating the context function are used. This improvement in classification accuracy is paid for by a substantial increase in computational requirements. An approximate algorithm, which cuts computational requirements by over one-half, is presented. Further reduction in computational requirements may be possible with a suggested hybrid algorithm. The search for an optimal implementation is furthered by an exploration of the relative merits of using spectral classes or information classes for classification and/or context function estimation. Finally, an unsuccessful attempt to devise a context measure for use in conjunction with context function estimation is described.

17. Key Words (Suggested by Author(s))	18. Distribution Statement
Remote sensing, classification algorithm, contextual classifier	

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified		

*For sale by the National Technical Information Service, Springfield, Virginia 22161

NASA - JSC

ACKNOWLEDGMENTS

The author wishes to acknowledge Professor Philip H. Swain for his numerous substantive and editorial suggestions that hav. added greatly to the clarity of this paper. An acknowledgement is also due to Professor Stephen B. Vardeman for suggestions that have added to the clarity and preciseness of the theoretical presentations. Funding for this research was provided by National Aeronautics and Space Administration Contract No. NAS9-15466.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Compound decision theory is employed to develop a general statistical model for classifying image data using spatial context. The classification algorithm developed from this model exploits the tendency of certain ground-cover classes to occur more frequently in some spatial contexts than in others. A key input to this contextual classifier is a quantitative characterization of this tendency: the context function. Several methods for estimating the context function are explored, and two complimentary methods are recommended. The contextual classifier is shown to produce substantial improvements in classification accuracy compared to the accuracy produced by a non-contextual uniform-priors maximum likelihood classifier when these methods of estimating the context function are used. This improvement in classification accuracy is paid for by a substantial increase in computational requirements. An approximate algorithm, which cuts computational requirements by over one-half, is presented. Further reduction in computational requirements may be possible with a suggested hybrid algorithm. The search for an optimal implementation is furthered by an exploration of the relative merits of using spectral classes or information classes for classification and/or context function estimation. Finally, an unsuccessful attempt to devise a context measure for use in conjunction with context function estimation is described. Recommendations for further research are included in the concluding chapter.

CHAPTER I - INTRODUCTION

The machine classification of multispectral image data collected by remote sensing devices aboard aircraft and spacecraft has usually been performed such that each pixel (picture element) is classified individually and independently [1]. The information used by this classifier is only spectral or, in some cases, spectral and temporal. There is no provision for using the spatial information inherent in the data. In contrast, when scanner data are displayed in image form, a human analyst routinely uses spatial information to establish a context for deciding what a particular pixel in the imagery might be. Using this context together with spectral information, the analyst may easily identify roads, delineate boundaries of agricultural fields, and differentiate between grass in an urban setting (e.g., lawns) and grass in an agricultural setting (e.g., pasture or forage crops) where a point-by-point classifier utilizing spectral information alone would have much difficulty in doing so.

The ECHO (Extraction and Classification of Homogeneous Objects) process is a variety of contextual classifier which has been found useful for classifying data sets which contain homogeneous objects that are large compared to the resolution of the imagery [2]. This classifier cannot be used effectively, however, if the data set does not contain a significant number of these large homogeneous objects.

A general statistical classification method which exploits both spatial and spectral information when classifying multispectral image data is the subject of this paper. This contextual classifier exploits the tendency alluded to earlier of certain ground-cover classes to be more likely to occur in some contexts than in others. In principle, this classifier can be used to advantage on any image data set, even those that do not have identifiable homogenous objects such as is generally the case in forested, urban and other inhomogeneous areas. However, the relatively high computational complexity of the contextual classifier limits its use to classification problems where the expected increase in accuracy is worth the increased computation cost.

The theoretical basis of this statistically based contextual classification algorithm is presented in Chapter II. This theoretical development is an elaboration and clarification of the development given by Swain and Vardeman in [3]. Chapter III presents exploratory experimental results including an evaluation of the performance of the algorithm on data which is simulated so as to meet the assumptions of the classification model and preliminary results of applying the algorithm to real Landsat data. Research problems indicated by these results are discussed at the end of Chapter III. The ensuing chapters discuss these research problems in detail.

CHAPTER II - THEORETICAL BASIS AND CLASSIFICATION MODEL

Consistent with the general characteristics of imaging systems for remote sensing, we assume a two-dimensional array of $N = N_1 \times N_2$ random observations X_{ij} having fixed but unknown classification ϑ_{ij}, as shown in Figure 1. The observation X_{ij} consists of n measurements (usually containing spectral and/or temporal information), while the classification ϑ_{ij} can be any one of m spectral or information classes* from the set $\Omega = \{\omega_1, \omega_2, \ldots, \omega_m\}$.

$$\begin{array}{|cccc|} \hline \vartheta_{11} & \vartheta_{12} & \cdots & \vartheta_{1N_2} \\ \vartheta_{21} & \vartheta_{22} & \cdots & \vartheta_{2N_2} \\ \vdots & & & \\ \vartheta_{N_1 1} & & \cdots & \vartheta_{N_1 N_2} \\ \hline \end{array}$$

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

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

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

* Spectral classes are spectrally differentiable subclasses of information classes (the classes of interest).

Similarly, let $\underline{\vartheta}$ be the vector of states (true classifications) associated with the observations in $\underline{X}$:

$$\underline{\vartheta} = [\vartheta_{ij} \mid i=1,2,\ldots,N_1; j=1,2,\ldots,N_2]^T.$$

Let the action (classification) taken with respect to pixel (i,j) be denoted by $a_{ij} \in \Omega$. The loss suffered by taking action a_{ij} when the true class is ϑ_{ij} is denoted by $\lambda(\vartheta_{ij}, a_{ij})$, for some fixed non-negative function $\lambda(\cdot,\cdot)$. In the most general case, the actions a_{ij} may be a function of all the observations in $\underline{X}$. For this case, the average loss suffered over the N classifications in the classification array is

$$L(\underline{\vartheta}, \underline{X}) = \frac{1}{N}\sum_{i,j}\lambda(\vartheta_{ij}, a_{ij}(\underline{X})).$$

The expected average loss (or risk) is then

$$R_{\underline{\vartheta}} = E\left[\frac{1}{N}\sum_{i,j}\lambda(\vartheta_{ij}, a_{ij}(\underline{X}))\right] \tag{1}$$

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

Our goal is to determine the dependence of the decision function $a_{ij}(\cdot)$ on $\underline{X}$ in such a way that, for any given classification array $\underline{\vartheta}$, the risk $R_{\underline{\vartheta}}$ will be minimum. One way to approach the problem of making $R_{\underline{\vartheta}}$ small is to view $\underline{\vartheta}$ as a realization of a random process in two dimensions and to derive a decision rule which is Bayes versus this "prior distribution" for $\underline{\vartheta}$. Simplifying assumptions concerning the nature of this process are generally made to find an associated Bayes rule which is both simple and has small $R_{\underline{\vartheta}}$ for most $\underline{\vartheta}$. This is the approach of Welch and Salter [4], 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.

Rather than looking for a prior distribution for $\underline{\vartheta}$ and an associated Bayes decision rule, we will adopt an approach for controlling $R_{\underline{\vartheta}}$ through $a_{ij}(\cdot)$ that is more closely related to the large body of statistical literature traceable to Robbins [5], and known as compound decision theory. See, for example, the works and references of Van Ryzin [6,7], and Vardeman [8].

The following notation will be useful. Let $\underline{\vartheta}^p \in \Omega^p$ and $\underline{X}^p \in (R^n)^p$ stand respectively for p-vectors of classes and n-dimensional measurements; each component of $\underline{\vartheta}^p$ is a variable which can take on any classification value; each component of $\underline{X}^p$ is a random n-dimensional vector which can take on values in the observation space.

Now we restrict the decision function $a_{ij}(\cdot)$ to depend only on a specified subset of the observations in $\underline{X}$. This subset includes, along with X_{ij}, p-1 observations spatially near to, but not necessarily adjacent to, X_{ij}. These p-1 observations serve as the spatial context for X_{ij} and are taken from the same spatial positions relative to pixel position (i,j) for all i and j. Call this arrangement of pixels together with X_{ij} the p-context array, several examples of which are shown in Figure 2. Group the p observations in the p-context array into a vector of observations $\underline{X}_{ij} = (X_1, X_2, \ldots, X_p)^T$ and let $\underline{\vartheta}_{ij}$ be the vector of true but unknown classifications associated with the observations in $\underline{X}_{ij}$. Note that the $\underline{\vartheta}_{ij}$ and $\underline{X}_{ij}$ are the particular instance of $\underline{\vartheta}^p$ and $\underline{X}^p$ associated with pixel position (i,j). Correspondence of the components of $\underline{\vartheta}_{ij}$, $\underline{X}_{ij}$, $\underline{\vartheta}^p$ and $\underline{X}^p$ to the positions in the p-context array is fixed but arbitrary except that the p^{th} components always correspond to the pixel being classified.

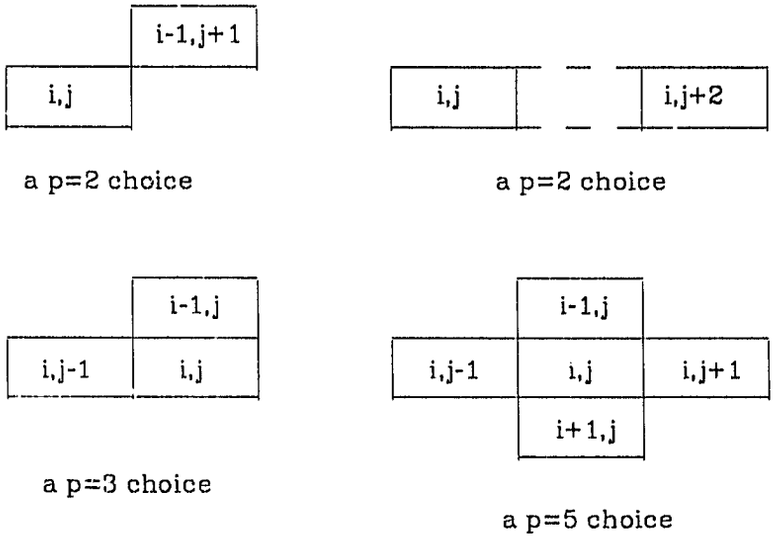


Figure 2. Examples of p-context arrays.

We shall seek an optimal decision rule of the form

$$a_{ij}(\underline{X}) = d(\underline{X}_{ij}) \tag{2}$$

for a fixed function $d(\cdot)$ mapping p-vectors of observations to actions. This decision rule is independent of location, depending only on the values of the observations in the p-context array and their relative locations. It provides the classification for the p^{th} pixel in the p-context array. The risk associated with any rule of this form is, from equation (1),

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

$$= \frac{1}{N}\sum_{\underline{\vartheta}^p \in \Omega^p}\ \sum_{\substack{i,j \text{ with} \\ \underline{\vartheta}_{ij}=\underline{\vartheta}^p}} E[\lambda(\vartheta_p, d(\underline{X}_{ij}))] \tag{3}$$

where ϑ_p is the p^{th} element of $\underline{\vartheta}^p$. If we require that the distribution of $\underline{X}$ is such that *every* $\underline{X}_{ij}$ for which $\underline{\vartheta}_{ij}=\underline{\vartheta}^p$ has the same marginal density, i.e., the marginal densities depend only on the measurement values in $\underline{X}_{ij}$ and the set of classifications in $\underline{\vartheta}_{ij}$ and not the location (i,j), we can then write

$$f_{ij}(\cdot \mid \underline{\vartheta}_{ij}=\underline{\vartheta}^p) = f(\cdot \mid \underline{\vartheta}^p). \tag{4}$$

Writing equation (3) in more detail using the class-conditional density $f(\cdot \mid \underline{\vartheta}^p)$, we have

$$R_{\underline{\vartheta}} = \sum_{\underline{\vartheta}^p \in \Omega^p} \frac{1}{N} \sum_{\substack{i,j \text{ with} \\ \underline{\vartheta}_{ij}=\underline{\vartheta}^p}} \int \lambda(\vartheta_p, d(\underline{X}^p)) f(\underline{X}^p \mid \underline{\vartheta}^p) d\underline{X}^p$$

$$= \sum_{\underline{\vartheta}^p \in \Omega^p} C(\underline{\vartheta}^p) \int \lambda(\vartheta_p, d(\underline{X}^p)) f(\underline{X}^p \mid \underline{\vartheta}^p) d\underline{X}^p$$

$$= \int \sum_{\underline{\vartheta}^p \in \Omega^p} C(\underline{\vartheta}^p) \lambda(\vartheta_p, d(\underline{X}^p)) f(\underline{X}^p \mid \underline{\vartheta}^p) d\underline{X}^p \tag{5}$$

where $C(\underline{\vartheta}^p)$, the "context function," is the relative frequency with which $\underline{\vartheta}^p$ occurs in the array $\underline{\vartheta}$. For any array $\underline{\vartheta}$, a decision rule $d(\underline{X}^p)$ minimizing $R_{\underline{\vartheta}}$ 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

$$\sum_{\underline{\vartheta}^p \in \Omega^p} C(\underline{\vartheta}^p) \lambda(\vartheta_p, a) f(\underline{X}_{ij} \mid \underline{\vartheta}^p). \tag{6}$$

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

$$\lambda(\vartheta,\alpha) = \begin{cases} 0, & \text{if } \vartheta = \alpha \\ 1, & \text{if } \vartheta \neq \alpha \end{cases} .$$

Then equation (6) simplifies and the decision rule becomes:

$$d(\underline{X}_{ij}) = \text{the action } \alpha \text{ which } \textit{maximizes}$$

$$\sum_{\substack{\underline{\vartheta}^p \in \Omega^p, \\ \vartheta_p = \alpha}} C(\underline{\vartheta}^p) f(\underline{X}_{ij} \mid \underline{\vartheta}^p). \tag{7}$$

A further assumption we make at this point is class-conditional independence of the observations (pixels) comprising $\underline{X}$. In this case,

$$f(\underline{X}_{ij} \mid \underline{\vartheta}^p) = \prod_{k=1}^{p} f(X_k \mid \vartheta_k) \tag{8}$$

where X_k and ϑ_k are the k^{th} elements of $\underline{X}_{ij}$ and $\underline{\vartheta}^p$, respectively. Evidence that this is a reasonable assumpution may be found in [9]. An approach for studying the effect of this assumption on this particular problem is also suggested in Chapter VIII. Invoking the class-conditional independence assumption, the decision rule (7) becomes:

$$d(\underline{X}_{ij}) = \text{the action } \alpha \text{ which maximizes}$$

$$\sum_{\substack{\underline{\vartheta}^p \in \Omega^p, \\ \vartheta_p = \alpha}} C(\underline{\vartheta}^p) \prod_{k=1}^{p} f(X_k \mid \vartheta_k) . \tag{9}$$

If the term $f(X_p \mid \alpha)$, corresponding to the pixel to be classified, is factored out of the sum the specific contribution due to context is made more apparent:

$$\left[\sum_{\substack{\underline{\vartheta}^p \in \Omega^p, \\ \vartheta_p = a}} C(\underline{\vartheta}^p) \prod_{k=1}^{p-1} f(X_k \mid \vartheta_k)\right] f(X_p \mid a) .$$

The context contribution is the term in brackets.

The optimal choice of $d(\cdot)$ cannot be implemented in practice since it depends on $C(\underline{\vartheta}^p)$ and the $f(X_k \mid \vartheta_k)$ which are unknown. Methods for estimating the $f(X_k \mid \vartheta_k)$ are well established from considerable experience in using the conventional non-contextual maximum likelihood decision rule [1]. When the classification set Ω consists of spectral classes, the $f(X_k \mid \vartheta_k)$ are assumed to be multivariate normal densities. In the case where the classification set Ω consists of information classes, the $f(X_k \mid \vartheta_k)$ are assumed to be weighted sums of multivariate normal densities.

Methods for estimating $C(\underline{\vartheta}^p)$ are not so well established as those for the $f(X_k \mid \vartheta_k)$. We can, however, expect that, at least for large $N = N_1 \times N_2$, a decision rule in which $C(\underline{\vartheta}^p)$ is replaced by an estimate $\hat{C}(\underline{\vartheta}^p)$ based on the $\underline{X}_{ij}$ will have risk $\hat{R}_{\underline{\vartheta}}$ appoximating that of the optimal rule. (We call this the "bootstrap effect.") That this is the case when p = 1 (equivalent to an optimal pointwise classifier with estimated *a priori* probabilities) and suitable forms of estimation are used is a consequence of the work of Van Ryzin [6]. The notion of attempting to approximate the risk of the best rule of the form shown in equation (2) for p > 1, given its first general treatment in Gilliland and Hannan [10], has not been as thoroughly studied as the p = 1 version. But related work for p > 1 in *sequence* versions of compound decision theory [11] suggests the validity of the generalization.

Comparing equation (6) with the results of Welch and Salter [4] and reinterpreting the $C(\underline{\vartheta}^p)$ as the marginal of an *a priori* distribution for $\underline{\vartheta}$, one

may view equation (6) as a generalization of the Welch and Salter contextual classification rule. The advantages of the present formulation are that one need make no possibly unrealistic assumptions about the distribution for $\underline{v}$ 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 immediately neighboring pixels would be useful, a possibility unique to this approach.

CHAPTER III - EXPLORATORY EXPERIMENTS AND DISCUSSION

The earliest experiments performed with the contextual classifier were exploratory in nature. The classifier concept feasibility was first established using simulated data, and the easiest and most obvious implementation of the contextual classifier was then used for a real Landsat data test. The test results from this implementation pointed to several research problems which are taken up in the following chapters.

Simulated Data Experiments

The initial experiments exploring the effectiveness of contextual classification using the set of discriminant functions defined by equation (9) to classify multispectral remote sensing data were performed on simulated data by Kit and Swain [12]. Simulated data were used so that the classification method's characteristics could be investigated undisturbed by unkown effects due to deviations of real data from the assumptions underlying the classifier. Each simulated data set was based on a non-contextual classification of multispectral remote sensing data which had been judged to be very accurate (produced by careful analysis of multitemporal data). Such a classification could be expected to embody the contextual content of the actual ground scene. Using the classification map and the associated estimated mean vectors and covariance matrices of the classes (developed in performing the non-contextual classification), data vectors were produced by a Guassian 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. We will refer to this template as the "reference classification."
2. All classes in the data set were known and represented.
3. All classes had multivariate Gaussian distributions with parameters typical of those found in real data.
4. All pixels were class-conditionally independent of adjacent pixels.
5. There were no mixture pixels.

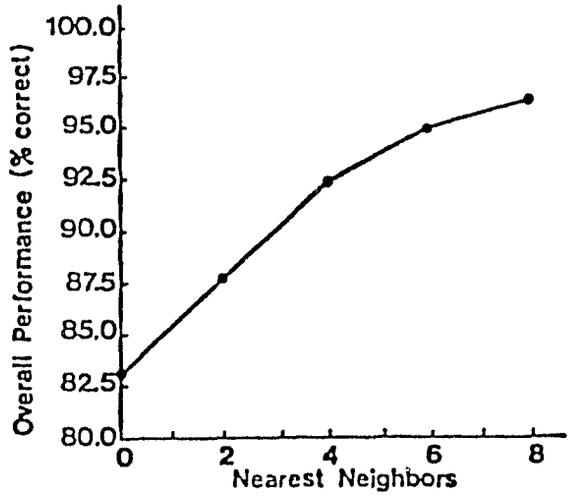
Data simulated in this manner are somewhat of an idealization of real remote sensing data, but the spatial organization of the simulated data is consistent with a real world scene and the overall characteristics of the data are consistent with the contextual classifier model. In essence, then, the experimental results based on the simulated data demonstrate the effectiveness of the contextual classifier, *given* that the underlying assumptions are satisfied. Experiments using the real data are discussed in the subsequent section and chapters.

Three classifications were selected and simulated data sets generated 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 of the simulated data sets was square, 50 pixels on a side.

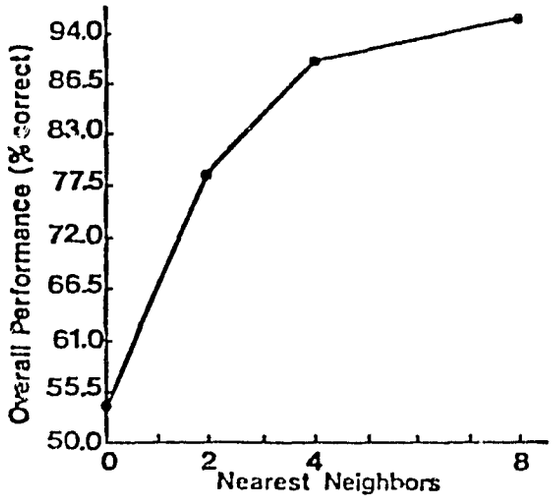
Figure 3 shows the classification results obtained. The "non-contextual" classification accuracy is plotted coincident with the vertical axis of each graph. Data set 1 was classified using successively 0, 2, 4, 6 and 8 neighboring pixels as context; data sets 2a and 2b were classified using 0, 2, 4 and 8 neighboring pixels. The accuracy improvement resulting from the use of contextual information in these simulated data sets was found to be quite significant.

As noted in Chapter II, to perform contextual classifications using the discriminant functions defined by equation (9), it is necessary to have available the class-conditional density functions for the classes to be recognized, $f(X_i \mid \vartheta_i)$, and the context function, $G(\underline{v}^p)$. In remote sensing applications, the class-conditional density functions are typically estimated from training samples. For the experiments described above, the $f(X_i \mid \vartheta_i)$ were taken to be the multivariate Gaussian distribitions from which the data were generated (these were originally the class-conditional density functions used to produce the reference classification used subsequently to produce the simulated data. An important question is how in practice to determine the context function. In the foregoing experiment, these relative frequencies were simply tabulated from the reference classification (actually, from an area somewhat larger than classified in this test). But in a real data situation, such a reference classification is not available, else there would be no need to perform any further classification.

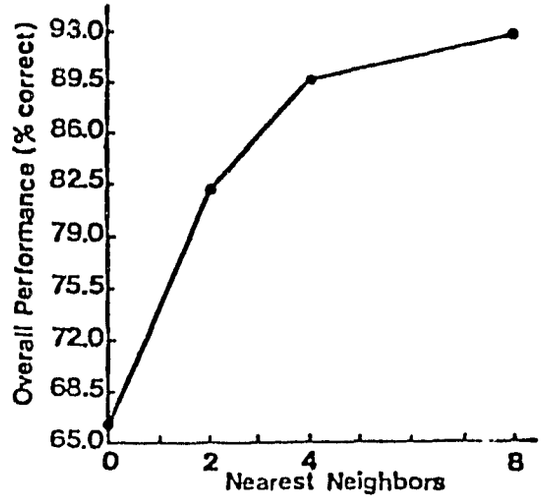
Looking towards extending the work of Kit and Swain to the real data case, we first investigated a straightforward approach to estimating the context function wherein we tabulated the relative frequencies from a uniform-priors non-contextual maximum likelihood classification of the same data.



(a)



(b)



(c)

Figure 3. Contextual classification of simulated data (from [12]): (a) data set 1; (b) data set 2a; (c) data set 2b.

Conceivably, one might then refine the estimate of the context function by making another estimate of the context function from the initial contextual classification, and even iterate in this way until no further improvements in classification accuracy were obtained. The crucial question here is how sensitive the contextual classification method is to the "goodness" of the context function estimate.

The potential of this iterative "classify-and-count" method was first tested on the simulated data set 2a. Prior to this test the classifications using context functions determined by tabulation from the reference classification were rerun using a tabulation of the context function from just the 50-pixel-square area classified, rather than from the larger area (276 × 320) used to obtain the estimate for the results presented in Figure 3. This was done to provide a better comparison to what could be accomplished using the iterative classify-and-count method. Also, the results were evaluated in terms of information classes rather than spectral classes, as was the case in Figure 3, in order to serve as a better comparison with real data tests.

Using the classify-and-count method, seven iterations (classifications followed by re-estimation of the context function) produced an improvement of 22.5 percent in overall accuracy compared to the non-contextual classification using equal *a priori* probabilities (from 70.5 percent to over 93 percent). Average-by-class accuracy rose only slightly (from 77.5 percent to 81 percent).* This compared with an increase of over 27.5 percent in overall

* 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 then 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.

accuracy (14.5 percent in average-by-class accuracy) obtained using the context function tabulated from the reference classification. These results are summarized in Figure 4.

As seen in Figure 4, several values of p (number of pixels in the p-context array) were used at each step of 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 the estimate of the context function for the next iteration. The best classification on the first iteration was obtained for p = 3 (nearest neighbors to the north and west), which was also the case for the second iteration. For the second iteration, the average-by-class accuarcy actually was slightly better for p=5 (four-nearest-neighbors), but the overall accuarcy was substantially higher for the p=3 choice. On the third iteration, the p=5 choice was selected since the overall accuracy was only slightly lower than for the p=3 choice while the average-by-class accuracy was substantially higher for the p=5 choice. The best classifications for the fourth and ensuing iterations were also the p=5 choice.

This implementation of the classify-and-count method involves a large 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 5, for simulated data set 2a the final result using this method was virtually the same as for the more involved procedure.

Just how much of the accuracy improvement was due to effectively making better estimates of the prior probabilities? After five iterations doing non-contextual classifications using prior probabilities estimated from the previous classification (the initial classification was a uniform-priors

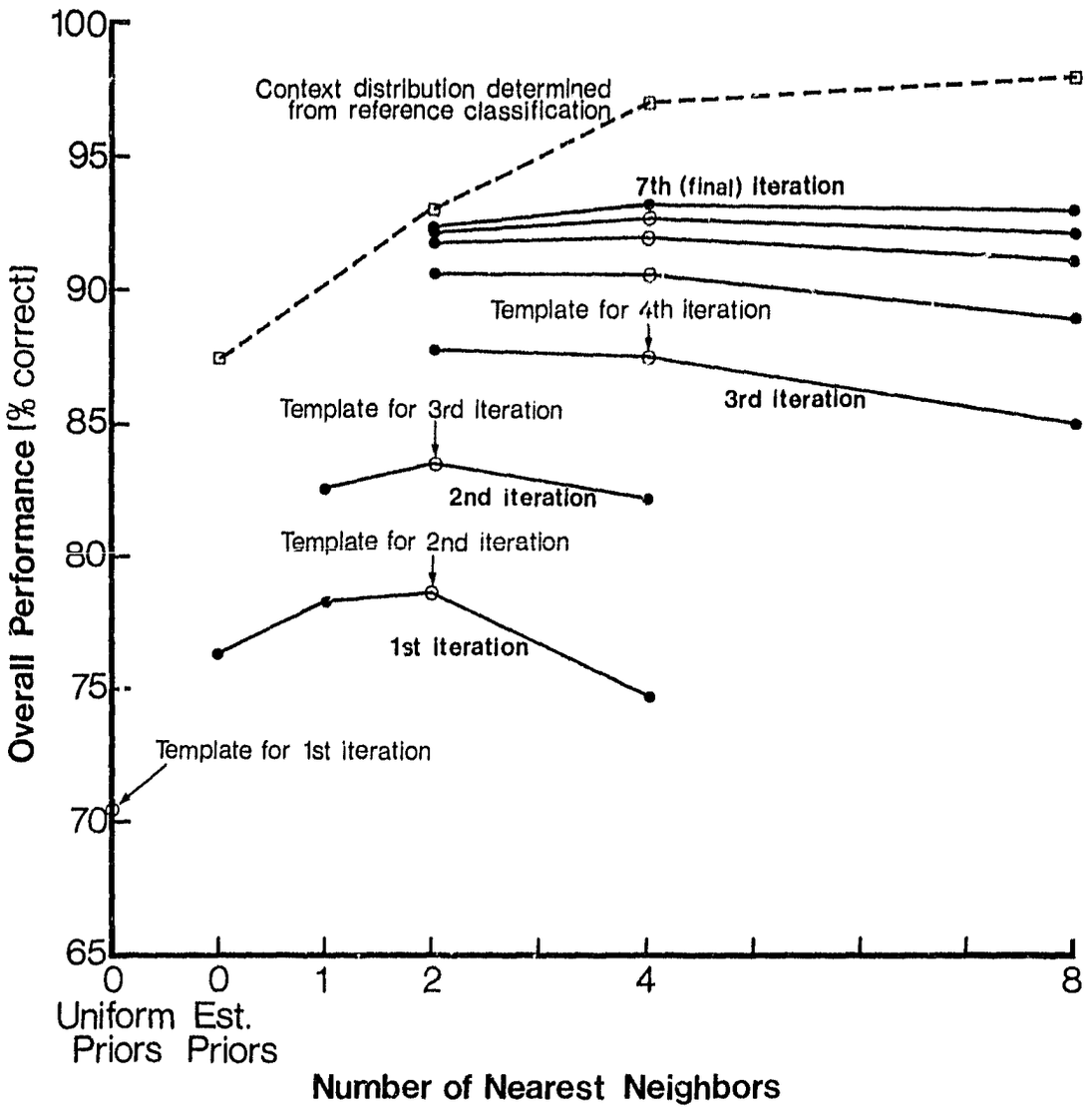


Figure 4. Contextual classification using the iterative classify-and-count method for estimating the context function (simulated data set 2a).

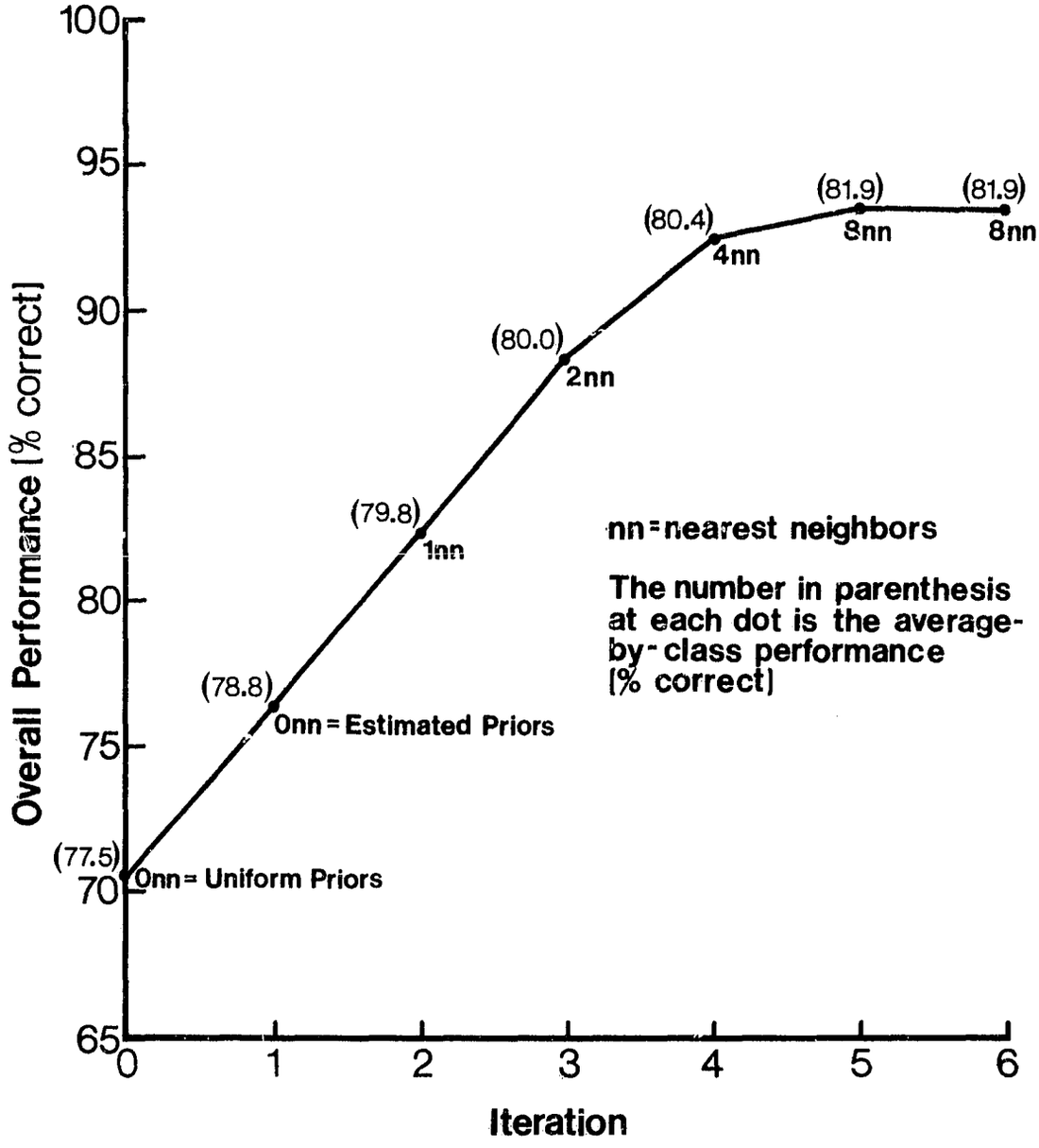


Figure 5. Contextual classification results based on simplified iterative technique (simulated data set 2a).

classification), the improvement in overall accuracy saturated at 87.1 percent, but the average-by-class accuracy had degraded to 64.7 percent. This compares closely to the non-contextual classification with prior probabilities tabulated from the reference classification, which had an overall accuracy of 87.5 percent and an average-by-class accuracy of 65.4 percent. It appears from this result that the context serves to improve the overall accuracy compared to that of the estimated-priors non-contextual result while resisting degradation in average-by-class accuracy.

Real Data (Landsat) 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 four-channel Landsat data was chosen which included approximately equal amounts of urban and agricultural area located to the southeast of Bloomington, Indiana. Parameters for the spectral classes were estimated using the 100-pixel-square area centered on the 50-pixel-square segment. A very careful non-contextual classification using 14 spectral classes was performed to delineate agricultural, urban and forested areas. As there were too few forest pixels to delineate forest test areas reliably, the classification was tested only for accuracy in discriminating between the agricultural and urban classes. 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 an aircraft on the same day as the Landsat pass (June 9, 1973).

The results obtained when using the full classify-and-count method on this data set were not as favorable as the results obtained with the simulated data. See Figure 6. The non-contextual 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 function 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 slighty poorer classification (84.2 percent overall, 83.8 percent average by class), although this difference may not be statistically significant. No apparent features in the scene would account for the difference (i.e., seen by visual inspection), but there is no reason to expect that Landsat data are strictly isotropic. This phenomenon will be pursued further in Chapter VII.

A second iteration was performed using the one-nearest-neighbor (north) classification from the first iteration as template for estimating the context function. Here the two-nearest-neighbor (neighbors to the north and west) classification was the best with an overall accuracy of 85.3 percent and average-by-class accuracy of 84.8 percent. Using the best second iteration result as template, the best classificaton for the third iteration was again the one-nearest-neighbor (north) case with 85.3 percent overall accuracy and 84.9 percent average-by-class accuracy. The fourth iteration produced no further improvement. The contextual classifier thus produced just over two percent improvement in both overall accuracy and average-by-class accuracy.

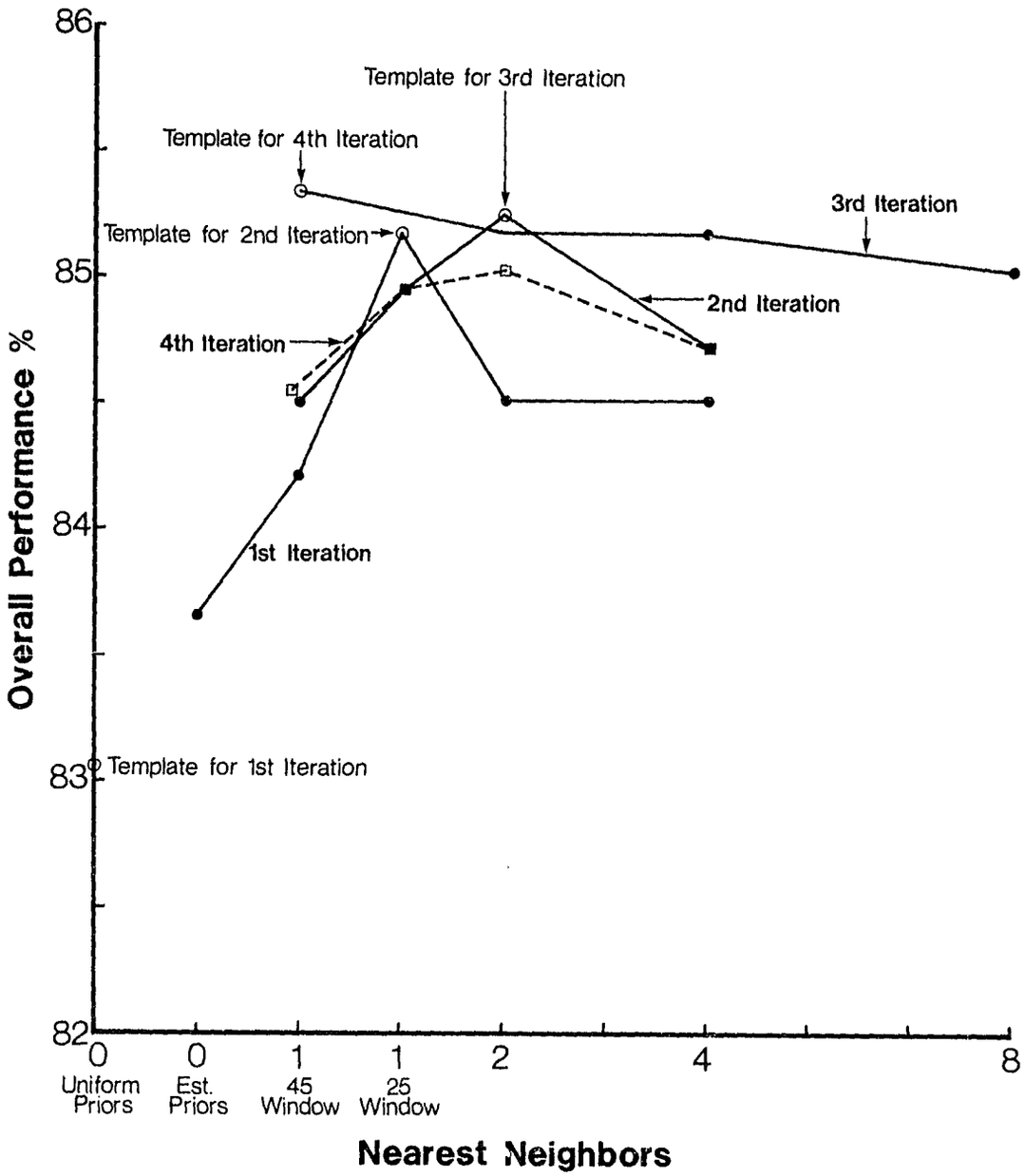


Figure 6. Contextual classification of the Bloomington, Indiana, data set using the classify-and-count method for estimating the context function. "25 window" refers to one-nearest-neighbor-to-the-north, "45 window" refers to one-nearest-neighbor-to-the-west.

The classify-and-count method was also tested on a 50-pixel-square agricultural scene. This was the northwest corner of the Large Area Crop Inventory Experiment (LACIE) Segment No. 1860 in Hodgeman County, Kansas. This data set was a four-channel Landsat data set collected on April 18, 1976. The class-conditional densities were estimated for the 16 spectral classes from randomly located training fields scattered throughout the entire 117-by-194-pixel Landsat data frame. The training fields were chosen by selecting pixel coordinates from a random number table and surrounding the selected pixel by the largest homogeneous rectangle up to field size 20-by-20. The classifications were tested for accuracy over five information classes (pasture, idle, wheat, corn and alfalfa) from "wall-to-wall" pixel-by-pixel ground truth.

The results obtained using this LACIE data set are summarized in Figure 7. Here the non-contextual classification using uniform prior probabilities had an overall accuacy of 78.7 percent and an average-by-class accuracy of 72.0 percent. The best classification (after five iterations) was a p=9 (eight-nearest-neighbors) classification with 80.5 percent overall accuracy and 73.0 average-by-class accuracy. Thus, the contextual classifier could only manage here a 1.8 percent improvement in overall accuracy and a 1.0 percent improvement in average-by-class accuracy.

Research Problems Indicated by the Exploratory Experiments

In the previous sections we saw that, on simulated data, the classify-and-count method produced estimates of the context function which in turn produced substantial improvements in classification accuracy. The classify-and-count method did not produce such good results with real Landsat data. It

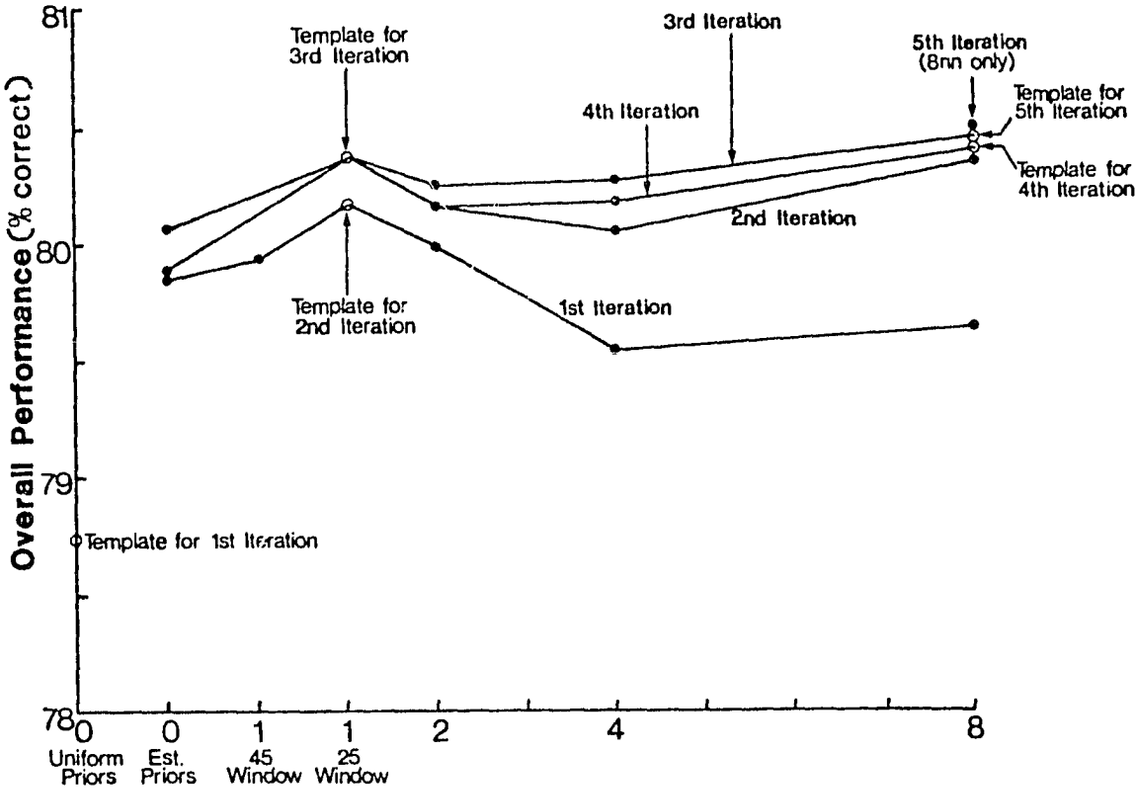


Figure 7. Contextual classification of LACIE Hodgeman County, Kansas, data set using the classify-and-count method for estimating the context function.

seems that for real data, the uniform-priors non-contextual classification is not a sufficiently accurate representation of the scene context to serve as basis for making a context function estimate which would lead to improved classification results. It may be that the classification of the simulated data was accurate enough because the class-conditional densities, $f(X_k \mid \vartheta_k)$, were modeled exactly, whereas the class-conditional densities were not modeled exactly for the real data classifications. The inaccuracy of the model in real data cases may contribute to producing estimates of the context function, $G(\underline{\vartheta}^p)$, which contain more erroneous class configuration counts than in the simulated data case. Such erroneous counts would cause poorer contextual classification results. Also, as we will see in Chapter IV, the classify-and-count method generally introduces a statistical bias into the context function estimate which would further contribute to the poor results observed. Whatever the reason for the poor performance of the classify-and-count method on real data, a better method for estimating the context function is needed. Chapter IV addresses this problem.

A second research problem area pointed out by the early experimental results is that a straightforward implementation of the contextual classifier is very computationally intensive. Depending on the number of neighbors used as context, the contextual classifier implemented on a PDP-11/45 computer needs anywhere from ½ hour to 6 hours elapsed time to classify a 50-pixel-square data set. Chapter V looks into strategies for reducing computational requirements.

A third research problem area involves certain assumptions which were made in the implementation of the contextual classifier used for the tests presented earlier in this chapter. First, the classification set Ω was assumed

to consist of spectral classes rather than information classes, and classifications were always made into spectral classes rather than information classes. This assumption is explored in Chapter VI. A second assumption was the class-conditional independence assumption represented by equation (8) in Chapter II. An approach for studying this assumption is discussed in Chapter VIII as a part of a discussion of areas for further research.

Chapters IV through VIII detail various approaches for dealing with these research problem areas. How these approaches relate to the main research problems and to our major goals of (a) advancing the theoretical understanding of this problem and (b) developing a contextual classification algorithm for use in practical problems is summarized in Figure 8. The solid lines represent connections of major significance, while the dotted lines represent less significant connections.

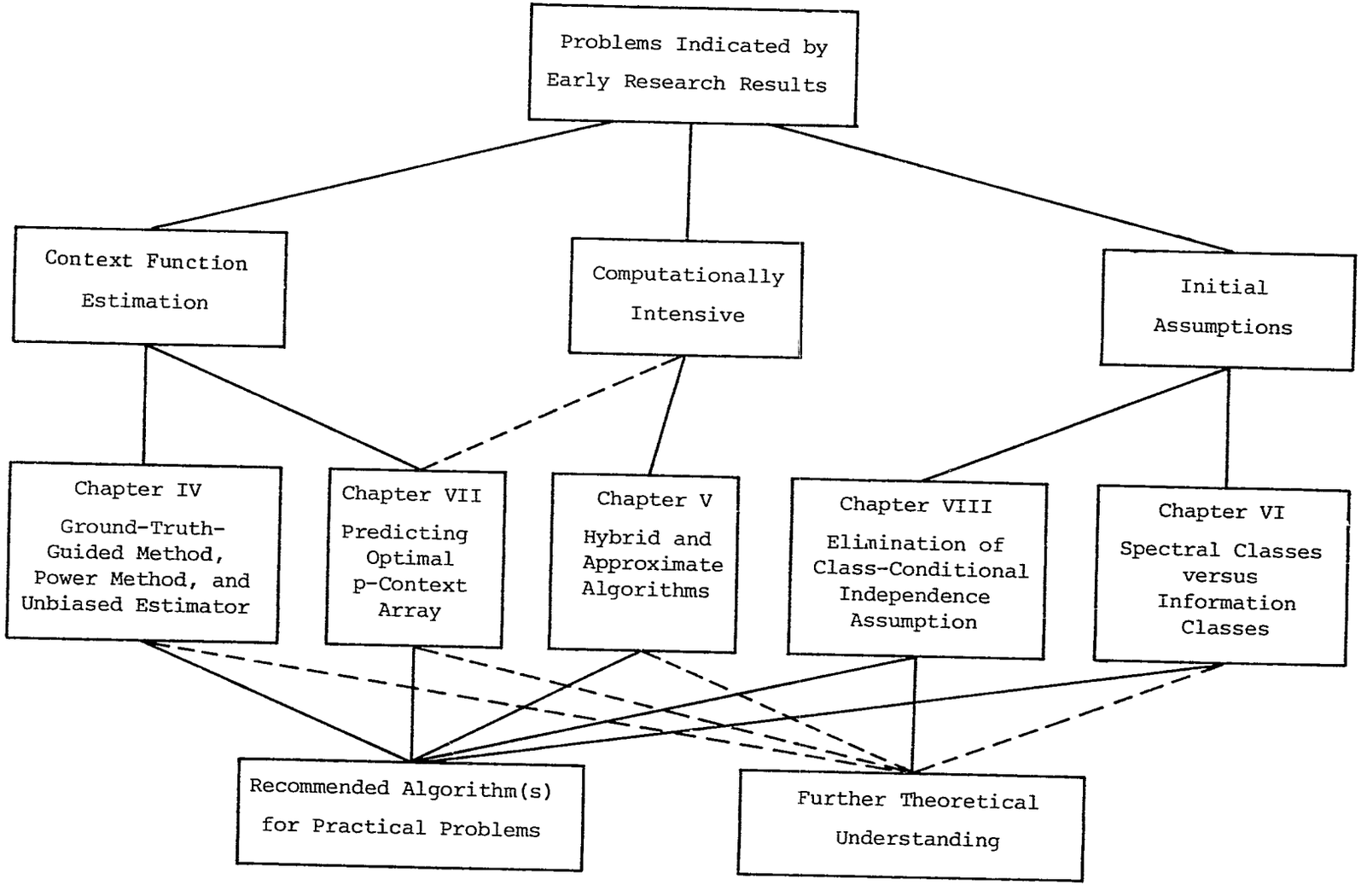


Figure 8. Interrelationships among topics of research.

CHAPTER IV - CONTEXT FUNCTION ESTIMATION

As we saw in Chapter III, the classify-and-count method of context function estimation produced unsatisfactory results for real Landsat data. These poor results spurred us to search for alternative methods of estimating the context function. Before we can discuss these alternative methods, however, we must briefly mention the spectral-class-versus-information-class question, since this question has some effect on the estimation methods to be discussed.

The contextual classifier implementation described in Chapter III performed classifications into spectral classes and used context functions taken over spectral classes. Information classes could have been used for either or both of these purposes. One could:

1. estimate the context function over spectral classes and classify into spectral classes (a pure spectral-class formulation), or
2. estimate the context function over spectral classes and classify into information classes, or
3. estimate the context function over information classes and classify into spectral classes, or
4. estimate the context function over information classes and classify into information classes (a pure information-class formulation).

These four options are explored in detail in Chapter VI. Having mentioned these implementation options, we can now turn to the search for effective context function estimation methods.

Ground-Truth-Guided Method

One alternative to the classify-and-count method is what we call the "ground-truth-guided method." The ground-truth-guided method is based on the idea that ground-truth information, if available, should improve the context function estimate when incorporated into the estimate. In this method, representative portions of the ground truth data are designated as a training set for estimating the context function and a test set for evaluating the classification results. The ground-truth data used for context function estimation must be in spatially contiguous blocks of size somewhat larger than the p-context array. The ground-truth data are, of course, represented in terms of information classes. When the estimation is to be done in terms of spectral classes rather than information classes, the following method is used:

1. Perform a non-contextual classification of the training set using uniform prior probabilities allowing the classifier to choose only among spectral classes associated with the information class designated by the ground truth.
2. Estimate the context function by tabulation from the resulting 100-percent accurate classification of the training set.
3. Classify the entire scene with the contextual classifier and evaluate the results over a test set disjoint from the training set.

When the estimation is to be done in terms of information classes, the restricted spectral class classification in step 1 above must still be performed. In this case, however, this classification is used to provide (by tabulation) an estimate of the weights for the weighted sums of normal densities that make up the class-conditional densities over information classes. The weights represent the relative frequency of observing a spectral class given that a particular information class was observed. The entire scene is then classified in terms of information classes using the contextual classifier, and evaluated over a test set disjoint from the training set, as in the spectral-class case.

Both the spectral- and information-class formulations (options 1 and 4) of the ground-truth-guided method were tested on two 50-pixel-square Landsat data sets. One data set was a LACIE data set from Hodgeman County, Kansas, containing pasture, wheat corn and fallow fields. This is the same data set described in Chapter III, except that two confounding spectral classes have been eliminated from the set Ω, leaving a total of 14 spectral classes. The other data set was from Tippecanoe County, Indiana, containing residential and commercial areas in northern Lafayette and West Lafayette as well as areas of forest, agriculture and water (the Wabash River). This data set was a four channel Landsat data set collected on June 20, 1976. Ground truth was obtained by visual inspection of large scale black and white aerial photographs taken on March 9, 1976 supplemented by ground inspection performed in January 1981. For both the Tippecanoe and LACIE data sets, the restricted spectral-class classification was performed over the first 25 lines of the data set and the context function was estimated over those 25 lines. Contextual classifications of the scenes were performed and classification

accuracies were evaluated over the last 25 lines as well as over the entire data set.

Tables 1 and 2 present the results from contextual classifications using four-nearest-neighbor (4nn) estimates of the context function (the p=5 choice in Figure 2) for both the spectral- and information-class formulations of the ground-truth-guided method (gtgm). These results are also compared to the accuracies obtained from uniform-priors and estimated-priors non-contextual maximum likelihood classifications. The prior probabilities for the estimated-priors non-contextual classifications were estimated by tabulation from the uniform-priors non-contextual classification. These results show that contextual classifications using the ground-truth-guided method for estimating the context function give significantly better results than non-contextual classifications on these data sets. For these cases, the spectral-class formulation of the ground-truth-guided method generally produces somewhat higher classification accuracies. However, since the spectral-class estimate of the context function has substantially more non-zero elements than the information-class estimate, contextual classifications using the spectral-class formulation generally take over twice the computer time required for the information-class formulation.

While this method produces estimates of the context function which give the best classification results of all methods discussed in this paper, it suffers the limitation that it requires large areas of spatially contiguous ground-truth data. When such detailed ground-truth data are not available, which is often the case since such ground truth is expensive and time-consuming to obtain, some other method is needed.

Table 1. Comparison of the contextual classifier using the ground-truth-guided method with non-contextual classifiers; Hodgeman County, Kansas, Landsat data set (14 spectral classes).

Classification	% Accuracy lines 26-50 Overall	lines 26-50 Average-by-Class	lines 1-50 Overall	lines 1-50 Average-by-Class
uniform priors	81.5	78.2	82.5	74.3
estimated priors	82.2	78.3	82.8	74.1
4nn gtgm, spectral	85.4	81.6	85.7	77.3
4nn gtgm, information	85.3	81.4	85.0	76.0

Table 2. Comparison of the contextual classifier using the ground-truth-guided method with non-contextual classifiers; Tippecanoe County, Indiana, Landsat data set.

Classification	% Accuracy lines 26-50 Overall	lines 26-50 Average-by-Class	lines 1-50 Overall	lines 1-50 Average-by-Class
uniform priors	82.7	81.7	81.8	83.4
estimated priors	84.2	82.0	83.7	83.7
4nn gtgm, spectral	88.7	91.1	89.3	90.7
4nn gtgm, information	88.2	87.3	88.2	86.2

Power Method

The classify-and-count method requires no ground-truth data besides that needed to estimate the class-conditional densities, $f(X_k \mid \vartheta_k)$. However, as we have seen earlier, this method does not produce consistently good estimates of the context function. In Chapter III we noted that the uniform-priors non-contextual classification does not seem to be a sufficiently accurate representation of the scene context for the classify-and-count method to perform well. The context function estimates generally contain several erroneous class configuration counts.

There are several ways in which the context function estimates from non-contextual classifications of real data could be "cleaned up." Assuming that the small relative frequency counts are more likely to be erroneous, one could employ a procedure which deletes all class configurations with frequency counts below a certain threshold. Or one could divide the count for each class configuration by a fixed number and take the integer part of the result as the new count, deleting all class configurations with counts that become zero.

Both of the aforementioned clean-up procedures could result in totally eliminating rarely occurring but valid classes from the context function. To avoid this problem, we devised an ad hoc procedure which we call the "power method."

The power method forms a new estimate of the context function by raising the relative frequency count for each class configuration to a power. For powers greater than one, the class configurations with larger counts are favored more heavily than those with relatively small (and possibly erroneous) counts. Conversely, for powers less than one, the class configurations

with large counts are not so heavily favored. At the extreme, a power of zero results in all class configurations being equally favored as in a uniform-priors non-contextual classification. In no case is an actually occurring class configuration deleted from the context function estimate.

The power method was first tested on a simulated data set to investigate the method's characteristics undisturbed by unknown effects due to inaccurate modeling of the real data sets. Spectral-class classifications using spectral-class context were performed using data set 2a (described in Chapter III). See Figure 9 for a summary of results. The results seem to indicate that when the model is exact, as the power is increased (up to a certain point), the classification results tend towards the results obtained when the context function is determined from the reference classification. Also, as expected, as the power used is decreased below unity, the results tend towards a uniform-priors non-contextual classification.

The power method was also tested on the Bloomington, Indiana, data set described in Chapter III using spectral-class context and classifications. Figure 10 summarizes the results using the power method on two-nearest-neighbors context (north and east neighbors) based on an estimate of $G(\underline{v}^p)$ from the non-contextual uniform-priors classification. Trading off overall accuracy against average-by-class accuracy, the best classification was produced using a power of 5, for which an overall accuracy of 87.0 percent and average-by-class accuracy of 86.1 percent was achieved. Note that the results in Figure 10 follow the same general trend as the simulated data results in Figure 9.

A second iteration of estimating $G(\underline{v}^p)$, this time over four-nearest-neighbors context, was then made based on the classifications listed in Figure

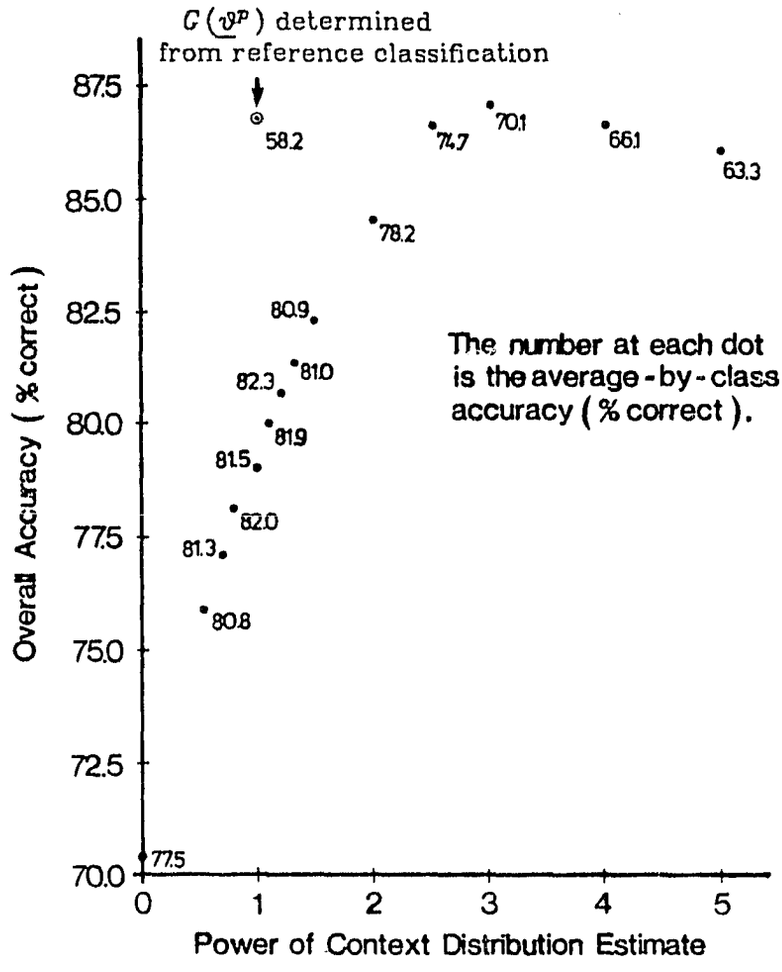


Figure 9. Power method results using as context one-nearest-neighbor (south) on the simulated data set. Context function, $C(\underline{v}^p)$, estimated from uniform-priors non-contextual classification except where noted otherwise.

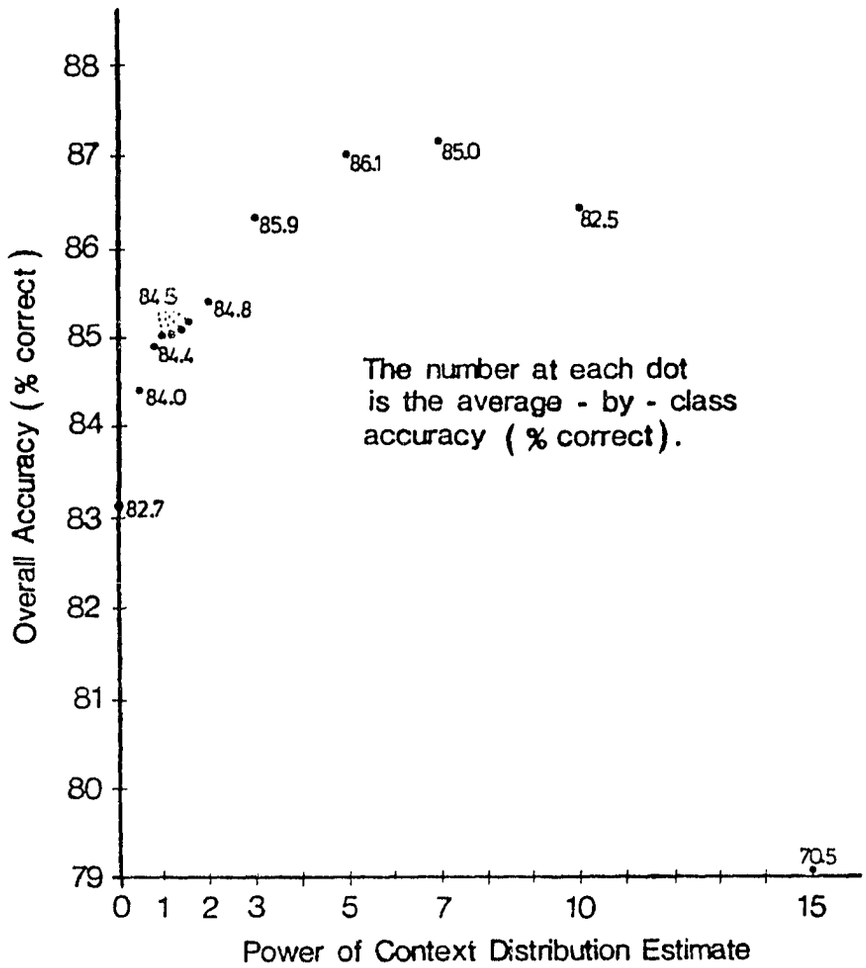


Figure 10. Power method results using two-nearest-neighbors (north and east) context on Bloomington, Indiana, data set. Context function, $G(\underline{v}^p)$, estimated from uniform-priors non-contextual classification.

10. The second estimate of $C(\underline{v}^p)$ based on the classification using the first estimate raised to a power of 10 produced the best classification results with an overall accuracy of 88.5 percent and an average-by-class accuracy of 87.5 percent (using $C(\underline{v}^p)$ raised to a power of 5). See Table 3 and Figure 11 for a summary of results. This second estimate of $C(\underline{v}^p)$ gave a total 5.4 percent improvement in overall accuracy and 4.8 percent improvement in average-by-class accuracy over the non-contextual classification. This compares with a 2.2 percent improvement in overall accuracy produced by the classify-and-count method in Chapter III.

Table 3. Second iteration power method results. Best four-nearest-neighbor classifications with $C(\underline{v}^p)$ based on the classifications in Figure 10.

Power Used in Figure 10	Power Used in this Classification	Accuracy, % Overall	Accuracy, % Average-by-Class
2	5	86.5	85.6
3	5	86.3	85.7
5	5	87.3	86.7
7	5	88.1	87.2
10	5	88.5	87.5
15	3	87.7	87.2

The power method was tested again on the Bloomington, Indiana data set, this time using information-class context and spectral-class classifications. (In implementing the power method elements of $C(\underline{v}^p)$ calculated from equation (33) in Chapter VI were raised to a power rather than elements of $H(\underline{\xi}^p)$.) Using a power of 7 in this case produced overall and average-by-class accuracies of 89.6 and 89.5 percent. These accuracies matched those produced in two iterations of the power method when spectral-class estimates of the

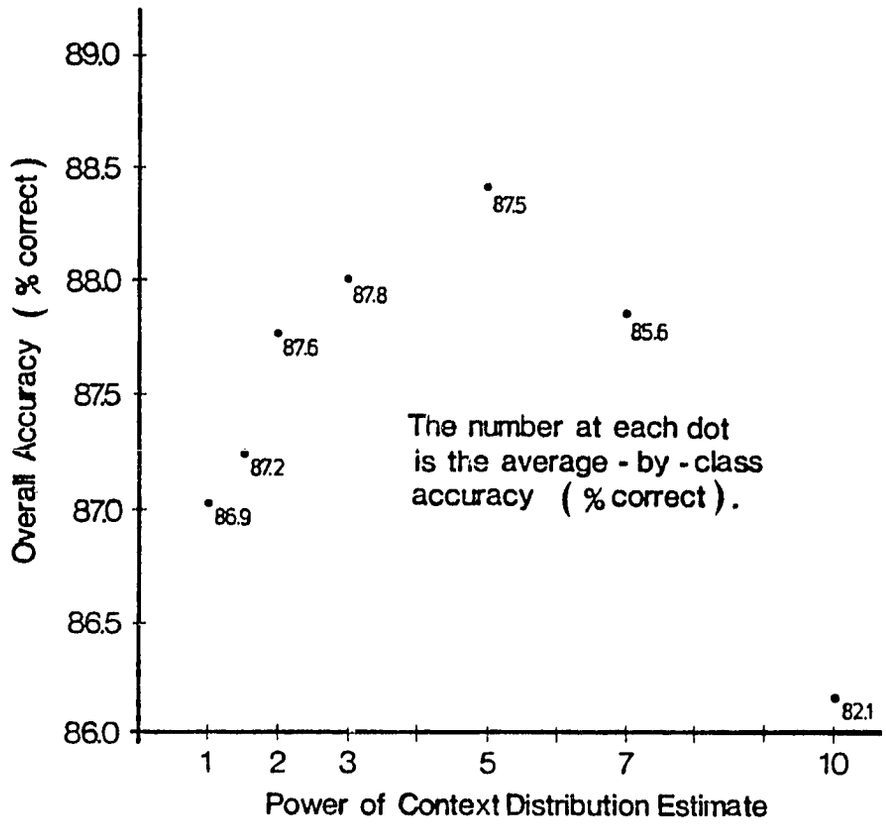


Figure 11. Power method results using four-nearest-neighbors context on Bloomington, Indiana, data set. Context function, $C(\underline{v}^p)$, estimated from two-nearest-neighbor (north and east) context classification with context function raised to power 10.

context function were used. Additional iterations in either case produced no further improvement in classification accuracies. Figure 12 compares using information-class estimates with using spectral-class estimates in the power method for the Bloomington, Indiana, data set.

A test of the power method was also performed on the LACIE data set (16 spectral classes) using spectral-class context and classifications. The spectral-class formulation results were similar to the Bloomington, Indiana, data set results. Again using two-nearest-neighbor context (neighbors to the east and west), the best classification was produced using a power of 7. Here the overall and average-by-class accuracies were 83.7 percent and 73.8 percent, respectively, as compared to overall and average-by-class accuracies of 78.7 and 72.0 percent, respectively, for the uniform-priors non-contextual case (evaluated over the entire scene). The best second-iteration result, using four-nearest-neighbor context, was produced with an estimate of $G(\underline{v}^p)$ made from the power of 15 first iteration classification and raised to a power of 10. This classification had an overall accuracy of 86.7 percent and average-by-class accuracy of 75.6 percent for an improvement of 8.0 percent and 3.6 percent, respectively, in overall and average-by-class accuracies. This compares to improvements of 1.8 percent and 1.0 percent, respectively, in overall and average-by-class accuracies produced by the spectral-class classify-and-count method when evaluated over the entire scene. When information-class context was used, the results were not as good. Two-nearest-neighbor context (north and west neighbors) raised to a power of 7 produced overall and average-by-class accuracies of 80.2 and 72.5 percent, respectively.

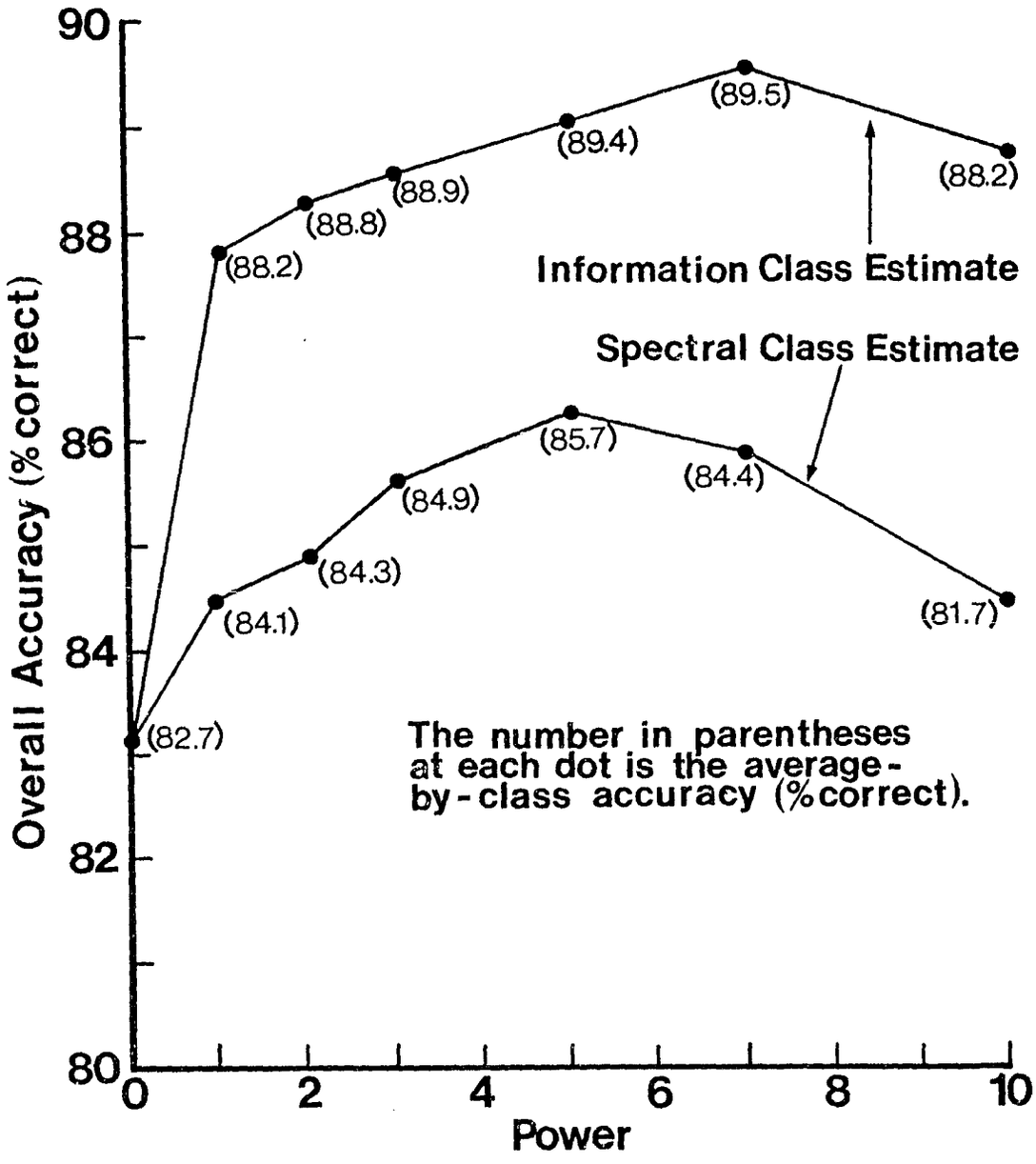


Figure 12. Summary of four-nearest-neighbor contextual classification results from the Bloomington, Indiana, data set. Here the power method is performed using both spectral-class and information-class estimates of the context function as tabulated from the uniform-priors non-contextual classification. Note that the power of zero result is equivalent to the uniform-priors non-contextual classification.

Prior to making the second-iteration estimates of $C(\underline{\vartheta}^p)$ in the above tests, it was assumed that a more accurate classification would necessarily produce a better estimate of $C(\underline{\vartheta}^p)$. The results quoted here indicate this is not always the case. This makes the power method more difficult to use, since classifications must be made using estimates of $C(\underline{\vartheta}^p)$ based on several classifications from the previous iteration in order to find the best estimate. Despite the good results possible with the power method, these ambiguities make this method difficult to use, and not useful for practical applications. A search for a better generally applicable method for estimating the context function has led to the unbiased estimation technique described next.

Unbiased Estimator

One tactic for seeking an optimal estimate of the context function, $C(\underline{\vartheta}^p)$, is to look for an estimator function, $T_{\underline{\vartheta}^p}(\underline{X})$, which minimizes the mean-squared error given by

$$MSE = E\left[T_{\underline{\vartheta}^p}(\underline{X}) - C(\underline{\vartheta}^p)\right]^2 . \tag{10}$$

Equation (10) can be rewritten as

$$MSE = Var[T_{\underline{\vartheta}^p}(\underline{X})] + b^2 \tag{11}$$

where $Var[T_{\underline{\vartheta}^p}(\underline{X})]$ is the variance of the estimate $T_{\underline{\vartheta}^p}(\underline{X})$ and b is the bias given by

$$b = E[T_{\underline{\vartheta}^p}(\underline{X})] - C(\underline{\vartheta}^p) . \tag{12}$$

Finding the minimum mean-squared-error estimate is generally a difficult task, but since bias represents a systematic error, a reasonable approach

would be to control bias before considering the variance. The best one can do in controlling bias is to seek an unbiased estimator, i. e., one for which $b = 0$.

As we saw in the previous section, the classify-and-count method performed poorly in tests on real Landsat data sets. One reason for this is that the estimate can be statistically biased. To prove this, consider the classification model as presented in Chapter II. In addition to the symbol definitions given there, we make the following definitions. Let $\underline{\hat{\vartheta}}$ be the vector of classifications

$$\underline{\hat{\vartheta}} = [\hat{\vartheta}_{ij} \,|\, i=1,2,\ldots,N_1; j=1,2,\ldots,N_2]^T$$

where $\hat{\vartheta}_{ij}$ is the classification estimate from a non-contextual classification of the observation X_{ij}. Let $\underline{\hat{\vartheta}}_{ij}$ be a p-vector of classification estimates associated with the observations in the p-context array $\underline{X}_{ij}$. Similarly, let $\underline{\hat{\vartheta}}^p$ be such an estimate associated with an arbitrary p-context array, $\underline{X}^p$. Let $\underline{\eta}^p \in \Omega^p$ represent an arbitrary p-vector of classes. The classify-and-count method can be described by the following estimator function for $C(\underline{\vartheta}^p)$:

$$T_{\underline{\vartheta}^p}(\underline{X}) \triangleq \hat{C}(\underline{\vartheta}^p) = \frac{1}{N}\sum_{i=1}^{N_1}\sum_{j=1}^{N_2} I(\underline{X}_{ij}, \underline{\vartheta}^p) \tag{13}$$

where

$$I(\underline{X}_{ij}, \underline{\vartheta}^p) = \begin{cases} 1, & \text{if } \underline{\vartheta}^p = \underline{\hat{\vartheta}}_{ij} \\ 0, & \text{otherwise.} \end{cases}$$

The expected value of $T_{\underline{\vartheta}^p}(\underline{X})$ is then

$$E[T_{\underline{\vartheta}^p}(\underline{X})] \triangleq E\left[\frac{1}{N}\sum_{i=1}^{N_1}\sum_{j=1}^{N_2} I(\underline{X}_{ij}, \underline{\vartheta}^p)\right] = \frac{1}{N}\sum_{i=1}^{N_1}\sum_{j=1}^{N_2} E\left[I(\underline{X}_{ij}, \underline{\vartheta}^p)\right]$$

$$= \frac{1}{N} \sum_{\underline{\eta}^p \in \Omega^p} \sum_{\substack{i,j \text{ with} \\ \underline{\vartheta}_{ij} = \underline{\eta}^p}} E\left[I(\underline{X}_{ij}, \underline{\vartheta}^p) \right] = \sum_{\underline{\eta}^p \in \Omega^p} G(\underline{\eta}^p) \int_{\substack{\underline{X}^p \in (R^n)^p \\ \text{with } \underline{\hat{\vartheta}}^p = \underline{\vartheta}^p}} f(\underline{X}^p \mid \underline{\vartheta}^p) d\underline{X}^p. \quad (14)$$

Equations (12) and (14) show that the bias of the classify-and-count method is the difference between a weighted sum of $G(\underline{\eta}^p)$ and $G(\underline{\vartheta}^p)$. Note that this bias is independent of N, and cannot be reduced by increasing the sample size. The bias can be non-zero or zero, depending of the values of $G(\underline{\eta}^p)$ and integrals in (14). To show this explicitly, let's consider the simple special case of a two-class problem (m=2) estimating non-contextual relative frequencies of classes (p=1) for univariate random observations (n=1). Let the non-contextual classifier used to produce $\underline{\hat{\vartheta}}$ be the uniform-priors maximum-likelihood classifier with the decision rule:

$$d(X_{ij}) = \text{the action } a \text{ which maximizes } f(X_{ij} \mid a)$$

for all $a \in \{\omega_1, \omega_2\}$. The densities, $f(X_{ij} \mid a)$, are assumed to be normal with mean and variance $\mu_1 = -1$ and ${\sigma_1}^2 = 1$ for class ω_1 and mean and variance $\mu_2 = 1$ and ${\sigma_2}^2 = 1$ for class ω_2. For class ω_1 we have:

$$E[T_{\omega_1}(X)] = \sum_{k=1}^{2} G(\omega_k) \int_{\substack{f(X \mid \omega_1) \\ \geq f(X \mid \omega_2)}} f(X \mid \omega_k)\, dX = \sum_{k=1}^{2} G(\omega_k) \int_{-\infty}^{0} f(X \mid \omega_k)\, dX$$

$$= G(\omega_1)\left[\frac{1}{2} + \text{erf}\frac{0-\mu_1}{{\sigma_1}^2}\right] + G(\omega_2)\left[\frac{1}{2} + \text{erf}\frac{0-\mu_2}{{\sigma_2}^2}\right]$$

$$= G(\omega_1)\left[\frac{1}{2} + \text{erf}\frac{0+1}{1}\right] + G(\omega_2)\left[\frac{1}{2} + \text{erf}\frac{0-1}{1}\right]$$

$$= .84G(\omega_1) + .16G(\omega_2). \quad (15)$$

The sum in (15) is equal to $G(\omega_1)$ only if $G(\omega_1) = G(\omega_2) = \frac{1}{2}$. For any other

values of $C(\omega_1)$ and $C(\omega_2)$ the estimate is biased. Similar comments apply for class ω_2 where we have

$$E[T_{\omega_2}(X)] = .16C(\omega_1) + .84C(\omega_2). \tag{16}$$

We have shown, then, that the classify-and-count method does indeed generally produce biased estimates of the context function.

The unbiased estimator we have adopted is presented in the statistical literature by Van Ryzin [6] and Hannan *et al.* [13]. This unbiased estimator can be most easily described by first considering the p=1 case and then generalizing to the arbitrary p-context array. For p=1, we examine the equation

$$\int h_k(X)\left[\sum_{l=1}^{m} f(X \mid \omega_l)C(\omega_l)\right]dX = \sum_{l=1}^{m}\left[\int h_k(X) f(X \mid \omega_l)dX\right]C(\omega_l) \tag{17}$$

where m is the number of classes; $f(X \mid \omega_l)$, $l=1,2,\ldots,m$, are the class-conditional densities described earlier; and the functions $h_k(X)$, $k=1,2,\ldots,m$, can be any set of m linearly independent functions. Equation (17) is valid provided all indicated sums and integrals are well defined, which will, for example, be the case when all of the functions in (17) are bounded. The functions $C(\omega_l)$ and $f(X \mid \omega_l)$ are always bounded because $C(\omega_l)$ is a relative frequency function and $f(X \mid \omega_l)$ is a multivariate normal density function. The functions $h_k(X)$ considered in the following development will also always be bounded.

The left-hand side of (17), which looks like the expected value of $h_k(X)$, can be estimated from the data $\underline{X}$ as follows:

$$\int h_k(X)\left[\sum_{l=1}^{m} f(X \mid \omega_l)C(\omega_l)\right]dX \cong \frac{1}{N}\sum_{i=1}^{N_1}\sum_{j=1}^{N_2} h_k(X_{ij}) \triangleq \overline{h}_k(\underline{X}) \tag{18}$$

where N, N_1 and N_2 are as defined in Figure 1, and $k \in \{1,2,\ldots,m\}$. Combining equations (17) and (18) we have

$$\overline{h}_k(\underline{X}) = \sum_{l=1}^{m} \left[\int h_k(X) f(X \mid \omega_l) dX\right] G(\omega_l) = \sum_{l=1}^{m} I_{kl}\, G(\omega_l) \tag{19}$$

where

$$I_{kl} \triangleq \int h_k(X) f(X \mid \omega_l)\, dX\,. \tag{20}$$

Applying (19) m times, once for each class, we can write

$$\begin{bmatrix} \overline{h}_1(\underline{X}) \\ \overline{h}_2(\underline{X}) \\ \vdots \\ \overline{h}_m(\underline{X}) \end{bmatrix} = \begin{bmatrix} I_{11} & I_{12} & \cdots & I_{1m} \\ I_{21} & I_{22} & \cdots & I_{2m} \\ \vdots & \vdots & & \vdots \\ I_{m1} & I_{m2} & \cdots & I_{mm} \end{bmatrix} \begin{bmatrix} G(\omega_1) \\ G(\omega_2) \\ \vdots \\ G(\omega_m) \end{bmatrix} \tag{21a}$$

This can be more succinctly represented in vector-matrix notation as

$$\underline{h} \triangleq I\, \underline{G}\,. \tag{21b}$$

Now $\underline{G}$ can be estimated by solving

$$\underline{G} \triangleq I^{-1}\, \underline{h} \triangleq \underline{T} \tag{22}$$

where $\underline{T} = (T_1(\underline{X}), T_2(\underline{X}), \ldots, T_m(\underline{X}))^T$ is the vector equivalent of $T(\underline{X})$ in (10), (11) and (12).

To show that $\underline{T}$ is indeed an unbiased estimator for $\underline{G}$, we note that

$$E(\underline{T}) = E(I^{-1}\, \underline{h}) = I^{-1}\, E(\underline{h})\,. \tag{23}$$

Looking at $E(\underline{h})$ element by element we have

$$E\left[\overline{h}_k(\underline{X})\right] \triangleq E\left[\frac{1}{N}\sum_{i=1}^{N_1}\sum_{j=1}^{N_2} h_k(X_{ij})\right] \tag{24a}$$

$$= \frac{1}{N}\sum_{i=1}^{N_1}\sum_{j=1}^{N_2} E\left[h_k(X_{ij})\right]$$

$$= \frac{1}{N}\sum_{i=1}^{N_1}\sum_{j=1}^{N_2}\int h_k(X_{ij})\, f(X_{ij}\,|\,\vartheta_{ij})\, dX_{ij}$$

$$= \frac{1}{N}\sum_{l=1}^{m}\sum_{\substack{i,j \\ with \\ \vartheta_{ij}=\omega_l}}\int h_k(X_{ij})\, f(X_{ij}\,|\,\vartheta_{ij})\, dX_{ij}$$

$$= \sum_{l=1}^{m} G(\omega_l)\int h_k(X)\, f(X\,|\,\omega_l)\, dX \tag{24b}$$

Thus

$$E(\underline{h}) = I\ \underline{G}$$

and (23) becomes

$$E(\underline{T}) = I^{-1}E(\underline{h}) = I^{-1}I\ \underline{G} = \underline{G} \tag{25}$$

proving that $\underline{T}$ is an unbiased estimator for $\underline{G}$.

It is convenient to use for the functions $h_k(X)$ a function of the class-conditional densities. More specifically, let $h_k(X) = (2\pi)^{\frac{n}{2}} f(X\,|\,\omega_k)$ and write (20) as

$$I_{kl} = (2\pi)^{\frac{n}{2}}\int f(X\,|\,\omega_k) f(X\,|\,\omega_l)\, dX$$

where n is the dimensionality of X. Assuming the ω_k are normally distributed spectral classes with respective mean vectors μ_k and covariance matrices Σ_k $(k=1,2,\ldots,m)$, we find

$$I_{kl} = \left[\det(\Sigma_k + \Sigma_l)\right]^{-\frac{1}{2}} \exp\{-\frac{1}{2}(\mu_k-\mu_l)^T (\Sigma_k+\Sigma_l)^{-1} (\mu_k-\mu_l)\}\,. \qquad (26)$$

When the ω_k are information classes, the I_{kl} are weighted sums of terms of the form given in (26). The weights are estimated by using the unbiased estimator with p=1 for the spectral classes which make up each information class being considered.

The calculation of the estimate of $\underline{G}$ can proceed in one of two alternative ways. The vector $\underline{h}$ can be calculated for the entire image (as in (21a)), then multiplied by I^{-1} to give $\underline{T} \cong \underline{G}$; or as the $h_k(X_{ij})$ are calculated at each data point (pixel), the product with I^{-1} can be performed. The average of these products over the entire image is then $\underline{T} \cong \underline{G}$. The methods are completely equivalent; the difference between them amounts to a change in order of summation. However, the second method must be used when this unbiased estimator is extended to the arbitrary p-context array case, because the use of the first method for large values of p would require an impractical amount of storage. In calculating the estimate of $G(\underline{v}^p)$ at each image data point using the second method, individual unbiased estimates of the prior probabilities of each class are made for each position in the p-context array, and cross-products of these prior probabilities are taken to form the unbiased estimate of $G(\underline{v}^p)$ based on that image point. To save computer storage space, the cross-products having values below a specified threshold are ignored. The estimate of $G(\underline{v}^p)$ for the entire image is the average of the

estimates of $C(\underline{y}^p)$ based on all the individual image points in the scene.

The unbiased estimator can be implemented so as to provide an adaptive estimate of the context function. The local context functio estimate for a particular $n_1 \times n_2$ block of image data is made from a $m_1 \times m_2$ block ($m_1 \geq n_1$ and $m_2 \geq n_2$). The $n_1 \times n_2$ block of image data is then classified using this local estimate of the context function. This process is repeated until the entire data set is classified. Better results have generally been obtained when $m_1 > n_1$ and $m_2 > n_2$. If $m_1 = n_1$ and $m_2 = n_2$, the context function estimate is not accurate for the pixels at the edges of the image data block being classified. Tests on three 50-pixel-square Landsat data sets have indicated good choices for n_1 and n_2 ranging from 10 up to 25 with the corresponding choices for m_1 and m_2 being 8 to 10 pixels larger than the values chosen for n_1 and n_2.

Table 4 presents the accuracies resulting from contextual classifications for three Landsat data sets using four-nearest-neighbor (4nn) estimates of the context function. The results using the spectral-class formulation are shown for the whole scene (non-adaptive) version and for an adaptive version employing local context function estimates for 25×25 pixel blocks made from the same 25×25 pixel block. The results using the information-class formulation are shown for an adaptive version employing estimates for various $n_1 \times n_2$ pixel blocks made from a $m_1 \times m_2$ pixel block centered on each $n_1 \times n_2$ pixel block. The uniform-priors non-contextual classification results are given for reference. The adaptive unbiased estimates generally performed best, especially when $m_1 > n_1$ and $m_2 > n_2$. The information-class formulation generally performed as well as the spectral-class formulation, with the information-class formulation performing substantially better on the Bloomington, Indiana, data set. As noted earlier in the discussion of the ground-truth-guided

Table 4. Comparison of the contextual classifier using various unbiased estimator formulations and the uniform-priors non-contextual classifier.

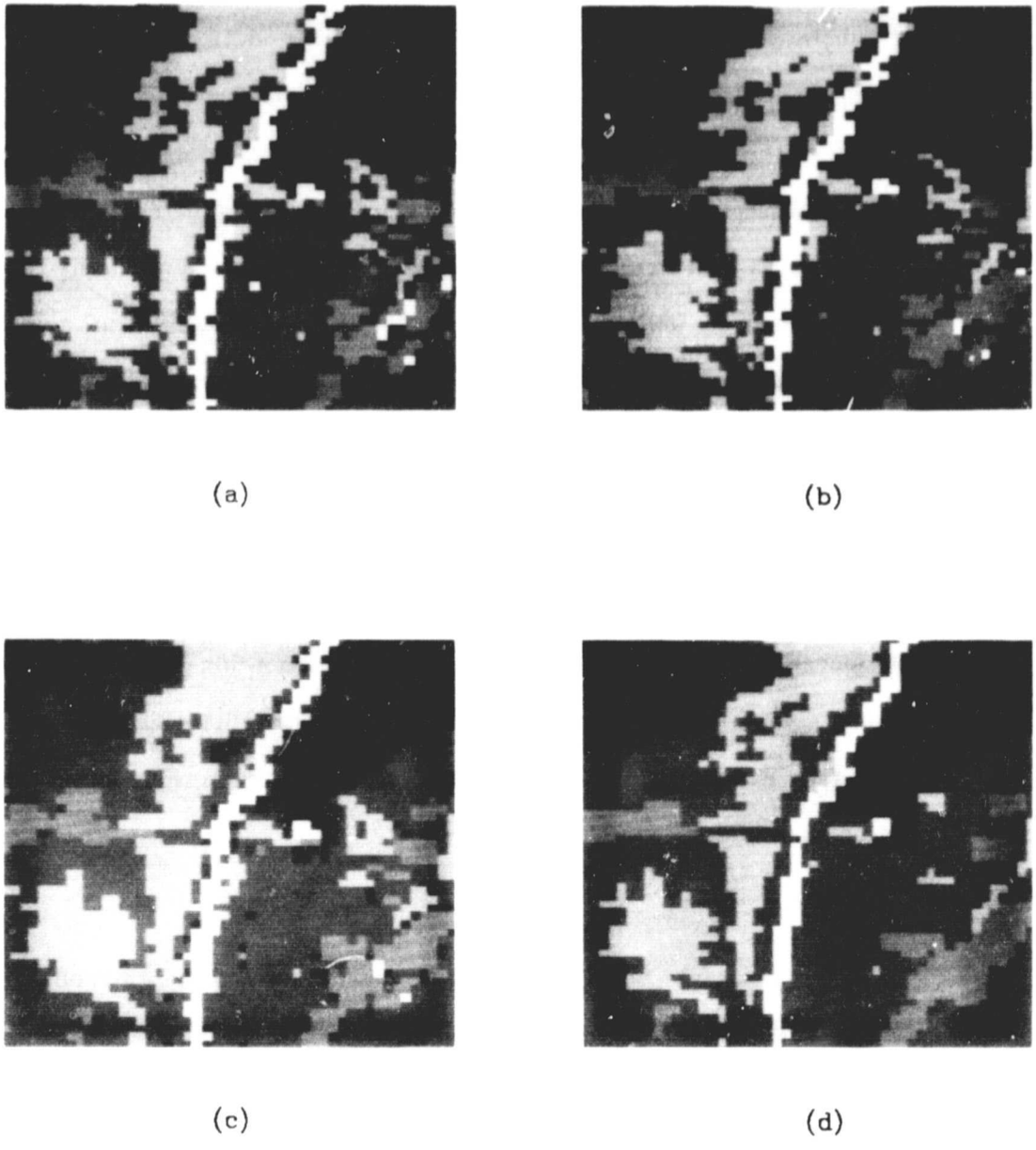
Data Set	Classification	%Accuracy Overall	%Accuracy Average-by-Class
Hodgeman County, Kansas, 50-pixel-square Landsat (evaluated over lines and columns 6 through 50; 14 spectral class LACIE)	uniform-priors non-contextual	82.0	75.9
	4nn unbiased, spectral class whole image est. (nonadaptive)	83.1	75.8
	4nn unbiased, spectral class adaptive est., 25×25 from 25×25	84.0	77.8
	4nn unbiased, information class adaptive est., 25×25 from 35×35	84.0	78.0
Bloomington, Indiana, 50-pixel-square Landsat	uniform-priors non-contextual	83.1	82.7
	4nn unbiased, spectral class whole image est. (nonadaptive)	84.4	84.4
	4nn unbiased, spectral class adaptive est., 25×25 from 25×25	84.3	83.9
	4nn unbiased, information class adaptive est., 17×17 from 25×25	88.9	88.3
Tippecanoe County, Indiana, 50-pixel-square Landsat	uniform-priors non-contextual	81.8	83.4
	4nn unbiased, spectral class whole image est. (nonadaptive)	86.2	87.9
	4nn unbiased, spectral class adaptive est., 25×25 from 25×25	86.7	88.1
	4nn unbiased, information class adaptive est., 25×25 from 25×25	86.2	89.1
	4nn unbiased, information class adaptive est., 10×10 from 20×20	86.9	89.7

method, the information-class formulation has the further advantage of having substantially fewer non-zero elements in the context function estimate, causing contextual classifications using an information-class formulation to require, in these tests, less than half the computer time required for contextual classifications using a corresponding spectral-class formulation.

Figure 13 shows computer generated gray-scale maps of classifications of the Tippecanoe County, Indiana, Landsat data set. The contextual classification looks visually closer to the reference classification than might be expected based on the accuracy improvement over the non-contextual classifications. This is due to the tendency of the contextual information here to provide a smoothing effect, making classification maps that are not only more accurate, but also more pleasing to the eye. This smoothing effect will not necessarily occur on all data sets. There is nothing inherent in the contextual classsification algorithm that would force smoothing when none is called for. The smoothing effect should only occur when the contextual information so indicates.

Summary

In our search to find successful methods for estimating the context function, we have explored the ground-truth-guided method, the power method, and a method utilizing an unbiased estimator. Tests on 50-pixel-square data sets have shown that all of these methods can provide estimates of the context function which produce contextual classifications with accuracies substantially higher than those obtained with a non-contextual classifier. We have seen, however, that the power method involves ambiguities (the optimal power value) that make it impractical for general use. Fortunately, the



(a) (b)

(c) (d)

Figure 13. Visual comparison of classification results, Tippecanoe County, Indiana, Landsat data set. (a) Uniform-priors non-contextual, (b) estimated-priors non-contextual, and (c) four-nearest-neighbor adaptive (17×17 from 27×27) unbiased estimator (d) reference classification.

unbiased estimator method performs excellently in those cases for which the power method would have been used, i.e., where large areas of spatially contiguous ground-truth are not available and hence the ground-truth-guided method cannot be employed.

The ground-truth-guided method can be used whenever large areas of spatially contiguous ground-truth data are available. In tests performed on 50-pixel-square data sets, the ground-truth-guided method outperformed the unbiased estimation method. However, the unbiased estimator produced contextual classifications which were nearly as accurate as those obtained using the ground-truth-guided method.

A pure spectral-class formulation was seen to perform slightly better than an information-class formulation for the ground-truth-guided method. An adaptive pure information-class formulation was seen to perform generally as well as or better than any other formulation of the unbiased estimator. In either case, the information-class formulation was seen to have a significant computational advantage.

The results of this chapter suggest candidates for successful implementations of the contextual classifier which should be tested with larger data sets. Further discussion of this topic will be deferred to Chapter VIII, after the other research areas mentioned in Chapter III are explored.

CHAPTER V - REDUCTION OF COMPUTATIONAL REQUIREMENTS

The contextual classification algorithm is very computationally intensive in both the spectral-class and information-class formulations, requiring a large amount of computer time. To reduce execution time, one could exploit the latest improvements in the raw speed of computer components and/or one could take advantage of special computer architectures involving multiple processing elements [14]. Alternative tactics explored in this chapter are (a) looking for a less computationally intensive algorithm which approximates the original contextual classification algorithm and (b) looking for a way to selectively apply the contextual classifier only where there is an advantage in doing so. We call the latter approach the "hybrid algorithm" because it uses a uniform-priors non-contextual classifier whenever that classifier can classify a given point "confidently," resorting to the contextual classifier only on "difficult" pixels. Before we consider the hybrid algorithm, we will first explore an algorithm which approximates the contextual classification algorithm as developed in Chapter II. If such an algorithm produces classifications that do not differ significantly in accuracy from the original algorithm, the approximate algorithm, possibly combined with the hybrid idea, would be the preferred algorithm in practical applications using conventional (serial) computers.

Approximate Algorithm

To come up with a reasonable approximate algorithm, one must examine the computer implementation of the original decision function*. Consider the case where the set Ω is defined over spectral classes, classification is into spectral classes, and the class-conditional independence assumption is taken. The densities $f(X_k \mid \vartheta_k)$ in equation (9) are assumed to be multivariate normal with mean vector M_{ϑ_k} and covariance matrix Σ_{ϑ_k} giving

$$f(X_k \mid \vartheta_k) = \left[\frac{1}{2\pi}\right]^{\frac{n}{2}} |\Sigma_{\vartheta_k}|^{-\frac{1}{2}} \exp\left[-\tfrac{1}{2}(X_k - M_{\vartheta_k})^T \Sigma_{\vartheta_k}^{-1}(X_k - M_{\vartheta_k})\right] \qquad (27)$$

where n is the dimensionality of the observation X_k (see [1] for the rationale behind this assumption in the non-contextual case). Using the multivariate normal assumption, the decision function in equation (9) becomes

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

where

$$d_a(\underline{X}_{ij}) = \sum_{\substack{\underline{\vartheta}^p \in \Omega^p, \\ \vartheta_p = a}} C(\underline{\vartheta}^p) \prod_{k=1}^{p} \left[\frac{1}{2\pi}\right]^{\frac{np}{2}} |\Sigma_{\vartheta_k}|^{-\frac{1}{2}} \exp\left[-\tfrac{1}{2}(X_k - M_{\vartheta_k})^T \Sigma_{\vartheta_k}^{-1}(X_k - M_{\vartheta_k})\right]. \qquad (28)$$

Let $d_a'(\underline{X}_{ij}) = \ln[d_a(\underline{X}_{ij}) \times (2\pi)^{\frac{pn}{2}}]$. Maximizing $d_a'(\underline{X}_{ij})$ is equivalent to maximizing $d_a(\underline{X}_{ij})$. Letting $Q_{\vartheta_k}(X_k) = (X_k - M_{\vartheta_k})^T \Sigma_{\vartheta_k}^{-1}(X_k - M_{\vartheta_k})$, we have

$$d_a'(\underline{X}_{ij}) = \ln\left[\sum_{\substack{\underline{\vartheta}^p \in \Omega^p, \\ \vartheta_p = a}} C(\underline{\vartheta}^p) \prod_{k=1}^{p} |\Sigma_{\vartheta_k}|^{-\frac{1}{2}} \exp[-\tfrac{1}{2} Q_{\vartheta_k}(X_k)]\right]$$

---

* For this study, the algorithm was implemented on a PDP-11/45 computer in the programming language "C". Test runs were also made on a PDP-11/70 computer.

$$= \ln\left[\sum_{\substack{\underline{\vartheta}^p \in \Omega^p, \\ \vartheta_p = a}} \exp\left[\ln C(\underline{\vartheta}^p) - \tfrac{1}{2}\sum_{k=1}^{p}\left[\ln|\Sigma_{\vartheta_k}| + Q_{\vartheta_k}(X_k)\right]\right]\right]$$

$$= \ln\left[\sum_{\substack{\underline{\vartheta}^p \in \Omega^p, \\ \vartheta_p = a}} \exp[F(\underline{X}_{ij}, \underline{\vartheta}^p)]\right] \tag{29}$$

where

$$F(\underline{X}_{ij}, \underline{\vartheta}^p) \triangleq \ln C(\underline{\vartheta}^p) - \tfrac{1}{2}\sum_{k=1}^{p}\left[\ln|\Sigma_{\vartheta_k}| + Q_{\vartheta_k}(X_k)\right].$$

In the simulated and real data sets studied (see Chapter III), the term $\exp[F(\underline{X}_{ij}, \underline{\vartheta}^p)]$ ranges over a larger negative exponential range than available on the PDP-11/45 (an exponential range of $10^{\pm 37}$ is available). To circumvent this problem it was necessary to use the following procedure.

Let

$$M_a(\underline{X}_{ij}) \triangleq \max_{\substack{\underline{\vartheta}^p \in \Omega^p, \\ \vartheta_p = a}} F(\underline{X}_{ij}, \underline{\vartheta}^p)$$

and rewrite $d_a'(\underline{X}_{ij})$ as follows:

$$d_a'(\underline{X}_{ij}) = \ln\left[\exp[M_a(\underline{X}_{ij})] \sum_{\substack{\underline{\vartheta}^p \in \Omega^p, \\ \vartheta_p = a}} \exp[F(\underline{X}_{ij}, \underline{\vartheta}^p) - M_a(\underline{X}_{ij})]\right]$$

$$= M_a(\underline{X}_{ij}) + \ln\left[\sum_{\substack{\underline{\vartheta}^p \in \Omega^p, \\ \vartheta_p = a}} \exp[F(\underline{X}_{ij}, \underline{\vartheta}^p) - M_a(\underline{X}_{ij})]\right]. \tag{30}$$

Calculating $d_a'(\underline{X}_{ij})$ in this way ensures that at least one term of the sum does not cause underflow because the exponential of the maximum term, $M_a(\underline{X}_{ij})$,

need not be calculated. This procedure also makes it less likely that other terms in the sum will cause underflow (the $F(\underline{X}_{ij}, \underline{\vartheta}^p)$ tend to be large negative numbers).

In checking out this particular implementation of the decision function, it was noted that $M_a(\underline{X}_{ij})$ was in most cases significantly larger than the logarithmic term in equation (30). This observation suggested the following approximation of the decision function:

$$d(\underline{X}_{ij}) = \text{the action } a \text{ which maximizes } M_a(\underline{X}_{ij}), \tag{31a}$$

or in the notation of equation (9):

$$d(\underline{X}_{ij}) = \text{the action } a \text{ which maximizes for all } \underline{\vartheta}^p \in \Omega^p \text{ with } \vartheta_p = a$$

$$G(\underline{\vartheta}^p) \prod_{k=1}^{p} f(X_k \mid \vartheta_k). \tag{31b}$$

Comparing equations (30) and (31a) one can see that the implementation of equation (31a) requires less computation and storage than equation (30). In equation (31a), the logarithmic term in equation (30) need not be calculated and the individual values of $F(\underline{X}_{ij}, \underline{\vartheta}^p)$ for a particular action a need not be stored; only the maximum value is needed. We would expect, then, that this approximate algorithm will take less computation time than the original algorithm for any data set. The effect of the approximation on classification accuracy, however, may be data dependent.

The performance of the approximate algorithm was compared with the original algorithm in tests using the simulated data set and the real data sets described in Chapter III. Included in the comparisons were algorithms that take only the three or five maximum terms in the summation in equation (9).

These additional algorithms serve to give an indication of how many terms in the summation are needed to produce classifications equivalent to those produced by the original algorithm. The results of this study are summarized in Table 5. The context function for the simulated data set test was estimated by tabulation from the reference classification from which the simulated data was generated and the context function for the LACIE data set was tabulated from the first 25 lines of a ground-truth-guided non-contextual classification as described in Chapter IV. (A ground-truth-guided classification is performed just like the usual non-contextual classification except that the classifier is restricted to selecting spectral classes from the information class indicated by the ground truth data.) Both data sets were evaluated over the entire 50-pixel square area. The context function for the Bloomington, Indiana, data set was tabulated from the entire 50-pixel square area of a ground-truth-guided non-contextual classification. Since the Bloomington data set has only 1317 ground-truth pixels, the ground-truth-guided classification degenerated to the usual unguided non-contextual classification over the remaining 1183 pixels. The Bloomington data set was evaluated over the 1317 ground-truth pixels. Eight-nearest-neighbor context was used in all cases.

As can be seen in Table 5, the approximate algorithm performed very well in terms of overall accuracy as compared to the original algorithm. The table also shows that in the two real data sets, the five largest terms of the sum in equation (9) are all that are needed to produce identical classifications to those produced by the full sum (the original algorithm).

The accuracy of the approximate algorithm was also tested in two cases where the "power method" was used for estimating the context function (see Chapter IV for a description of the power method). Table 6 displays the

classification accuracies resulting from applying the power method to the Bloomington and LACIE data sets in the same manner as described in Chapter IV.

Table 5. Performance of approximate algorithm in terms of accuracy. Context function estimated from ground-truth-guided classification.

Data Set	Overall Accuracy, %			
	Orig. Alg., Eq. (9)	5 Largest Terms of Sum in Eq. (9)	3 Largest Terms of Sum in Eq. (9)	Approx. Alg., Eq. (31a&b)
Simulated	96.84	96.88	97.04	97.04
LACIE	87.52	87.52	87.52	87.47
Bloomington	95.60	95.60	95.52	95.52

Table 6. Performance of approximate algorithm in terms of accuracy. Context function estimated using power method.

Data Set	Overall Accuracy, %	
	Original Algorithm, Equation (9)	Approximate Algorithm, Equation (18a&b)
Bloomington	88.46	88.38
LACIE	86.70	86.66

Again the approximate algorithm produced overall accuracies that were very close to those produced by the original algorithm. To put these minor accuracy differences in proper perspective, it helps to note that a conventional uniform-priors non-contextual classifier produced overall accuracies of 83.07 percent on the Bloomington data set and 78.73 percent on the LACIE data set.

The approximate algorithm was compared with the original algorithm in terms of computation time on the simulated data set and the two real Landsat data sets. Highly optimized versions of each algorithm (written in the "C" programming language) were run on PDP-11/45 and PDP-11/70 computers. Also compared to these two algorithms was a highly optimized version of the original algorithm that simply ignored underflows rather than attempting to circumvent them. This version allowed comparison of the approximate algorithm to a simulated implementation of the original algorithm on a computer with adequate exponential range.

The length of time the classifier took to process the 50-pixel square data sets depended strongly on the number of nonzero elements of the context function. (The number of terms that need to be evaluated in the sum in equation (9) and the number of terms to be compared in the maximization of equation (31b) is equal to the number of nonzero elements in the context function.) The ratio of timings between the three programs remained fairly consistent, however, across all data sets. Tables 7 and 8 display typical quiet system* timings on a PDP-11/45 computer for cases of few nonzero elements of the context function (480) and relatively large number of nonzero elements (2193). Table 9 gives the timings for the case displayed in Table 8, but run on a PDP-11/70 computer.

The three tables show that the approximate algorithm averaged less than half the real or user time taken by either of the other two algorithms. This amounts to a significant improvement in computation time.

---

* The runs were made during early morning hours when few other tasks were being performed by the computer.

Table 7. Performance of approximate algorithm in terms of timings. 50-pixel-square LACIE data set, two-nearest-neighbor context, 480 nonzero elements in context function, PDP-11/45 computer.

Algorithm	Time in Seconds*
Original Algorithm With Underflow Protection	2636
Original Algorithm Without Underflow Protection	2388
Approximate Algorithm	1185

Table 8. Performance of approximate algorithm in terms of timings. 50-pixel-square simulated data set, two-nearest-neighbor context, 2193 nonzero elements in context function, PDP-11/45 computer.

Algorithm	Time in Seconds*
Original Algorithm With Underflow Protection	14702
Original Algorithm Without Underflow Protection	14290
Approximate Algorithm	8675

* Timings are given in terms of "user time", which is essentially time spent doing computations.

Table 9. Performance of approximate algorithm in terms of timings. 50-pixel square simulated data set, two-nearest-neighbor context, 2193 nonzero elements in context function, PDP-11/70 computer.

Algorithm	Time in Seconds
Original Algorithm With Underflow Protection	5832
Original Algorithm Without Underflow Protection	6573
Approximate Algorithm	2526

In summary, experimental results from one simulated and two real data sets show that on these data sets the approximate algorithm takes significantly less computer time while producing classifications that do not differ significantly in accuracy from classifications produced by the original algorithm. By the nature of the approximate algorithm, it is expected that similar time savings will occur when the approximate algorithm is used on other data sets. Whether or not the accuracy results presented here can be expected with other data sets depends on the extent to which the data sets tested here are representative of remotely sensed data in general. We feel that they are fairly representative.

Hybrid Algorithm

A second way to produce classifications with accuracy comparable to the original contextual classification algorthm but with less computation may be to use a "hybrid" algorithm which would use a uniform-priors non-contextual classifier whenever that classifier can classify a given point "confidently," resorting to the contextual classifier only on "difficult" pixels. In other words, when the multispectral information alone at a given pixel were adequate to

confidently classify the pixel, the contextual information would not be used.

A simple measure of the "confidence" of classification by a uniform-priors non-contextual classifier would be the magnitude of the largest discriminant function at a given pixel. Another measure would be the difference between the classifier's two largest discriminant function values at a given pixel divided by the largest discriminant function ("normalized difference"). If either of these factors exceeded specified thresholds, the classification indicated by the uniform-priors non-contextual classifier would be accepted. Otherwise, the contextual classifier would be invoked. Such a method should save considerable computation time, depending on the percentage of pixels that must be classified by the contextual classifier. Classification accuracy should not suffer significantly because the pixels classified "confidently" by the uniform-priors non-contextual classifier presumably would have been classified identically by the contextual classifier.

A confidence measure must be efficient and accurate in order to be used to good advantage here. A perfectly efficient and accurate confidence measure for this problem would indicate (or flag) a low confidence classification if and only if the non-contextual classification would be different than the contextual classification. A practical confidence measure could approach the accuracy ideal of flagging all pixels that have different non-contextual classifications from the contextual classification. Such a practical confidence measure could not be expected to be perfectly efficient, however, for any confidence measure would be expected to produce a number of false alarms (pixels being flagged which have identical non-contextual and contextual classifications) since we would expect by chance that a portion of the low confidence non-contextual classifications will have the same classification as

the contextual classification. An efficient and accurate confidence measure would flag all or nearly all the pixels that had different non-contextual and contextual classifications, and would also produce a minimum number of false alarms.

A preliminary test of the hybrid approach was performed using the 50-pixel-square Tippecanoe County, Indiana, data set. In this test, the contextual classification compared with the uniform-priors non-contextual classification used a four-nearest-neighbor context function estimated by using the pure information-class formulation of the adaptive unbiased estimator of context (Chapter IV). The best result, in terms of efficiency and accuracy, was obtained by flagging those pixels which were below a threshold value of .90 for the normalized difference or below a threshold of 10^{-3} for the largest discriminant function. Here 756 pixels were flagged (out of 2500 in the image), 621 of which were false alarms. There were 287 pixels which were actually different between the contextual and non-contextual classifications. Thus, 149 pixels that should have been flagged were not flagged. The non-contextual classification had an overall accuracy of 81.8 percent and average-by-class accuracy of 83.4 percent. The contextual classification had overall and average-by-class accuracies of 86.9 and 89.7 percent, respectively. The hybrid classification had overall and average-by-class accuracies of 84.0 ad 86.6 percent, respectively.

The results indicate that these simple confidence measures are not very accurate or efficient indicators of pixels that would be classified differently by the non-contextual and contextual classifiers. It is apparent that a more sophisticated approach is needed. Such an approach would take into account the location of each measurement in the measurement space in relation to

the multidimensional contours of the class-conditional density functions. A confidence (or reliability) measure of this type is suggested in Alvo and Goldberg [15], but will not be pursued further here.

CHAPTER VI - SPECTRAL CLASSES VERSUS INFORMATION CLASSES

In Chapter IV we briefly mentioned the spectral-class-versus-information-class question. This chapter addresses this question in detail. To reiterate, the spectral-class-versus-information-class question involves four different options. One could:

(1) estimate the context function over spectral classes and classify into spectral classes (a pure spectral-class formulation), or

(2) estimate the context function over spectral classes and classify into information classes, or

(3) estimate the context function over information classes and classify into spectral classes, or

(4) estimate the context function over information classes and classify into information classes (a pure information-class formulation).

The question is, which option is the best to use?

In Chapter IV we concluded that a pure spectral-class formulation performed slightly better than an information-class formulation for the ground-truth-guided method. A pure information-class formulation generally performed as well as or better than any other formulation of the unbiased estimator. In either case we noted that the pure information-class formulation had a significant computational advantage over the spectral-class formulation. This chapter explores the spectral-class-versus-information-class question with respect to the simplest context function estimation method: the

classify-and-count method. The tests of the classify-and-count method described in Chapter III assumed spectral-class context and spectral-class classification (option 1). We will now discuss spectral-class context and information-class classification (option 2).

Spectral-Class Context and Information-Class Classification

Since classification results are normally evaluated over information classes rather than spectral classes, it may prove fruitfull to classify directly into information classes. When a classification problem is formulated so as to classify into spectral classes, one is actually maximizing accuracy with respect to spectral classes rather than information classes. In order to maximize accuracy with respect to information classes, one must formulate the classification problem so as to classify into information classes. In spite of this theoretical justification for classifying into information classes, it has generally been noted in non-contextual classification problems that information-class classification does not always produce an improvement in classification accuracy over that produced by a spectral-class classification. Hixson *et al.* [16] could only cautiously report a small improvement in classification accuracy in certain cases where a non-contextual maximum likelihood classification was done directly into information classes rather than into spectral classes. Will information-class classification fulfill its theoretical promise for the contextual-classifier when utilizing spectral-class context?

The contextual classification decision rule must be reformulated slightly to study this question. Let the set $\Omega=\{\omega_1,\omega_2,\ldots,\omega_m\}$ represent spectral classes and the set $\Gamma=\{\gamma_1,\gamma_2,\ldots,\gamma_n\}$, $n\leq m$, represent information classes. Note that

each element of Γ is a subset of the spectral classes such that if $\omega_i \in \gamma_j$ then $\omega_i \notin \gamma_k$ for k $\neq$ j and $\bigcup_{j=1}^{n} \gamma_j = \Gamma$. Let $\underline{\vartheta}^p \in \Omega^p$ and $\underline{\xi}^p \in \Gamma^p$ stand for p-vectors of classes over spectral and information classes, respectively.

Where the possible actions are defined over information classes, and the contextual information is defined in terms of spectral classes, the decision rule is obtained by maximizing a function as in equation (7) summed over the spectral classes contained in the action (information class) considered. Invoking the class-conditional independence assumption as in equation (9), the decision rule becomes:

$$d(\underline{X}_{ij}) = \text{the action } \alpha \in \Gamma \text{ which maximizes}$$

$$\sum_{\sigma \in \alpha} \left[\sum_{\substack{\underline{\vartheta}^p \in \Omega^p, \\ \vartheta_p = \sigma}} C(\underline{\vartheta}^p) \prod_{k=1}^{p} f(X_k \mid \vartheta_k) \right] \tag{32}$$

where the σ are the spectral classes making up information class α, and ϑ_k and X_k are the k^{th} elements of $\underline{\vartheta}^p$ and $\underline{X}_{ij}$, respectively. Note that this classification decision rule entails no more computation than a pure spectral-class decision rule as in equation (9). In fact, slightly less computation is needed with this decision rule because fewer comparisons are needed between values for $d(\cdot)$ since there are fewer possible actions α when classification is done into information classes.

This decision rule was tested on simulated data set 2a. The results are reported in Table 10. Here the context function was tabulated from the original reference classification. In all cases, except the uniform-priors non-contextual classification, the information-class classification gave results

which were virtually identical to the spectral-class classification. The information-class classification was more accurate than the spectral-class classification for the uniform-priors non-contextual case. These results would seem to indicate that the potential of contextual classification into information classes using spectral-class context is limited in terms of accuracy improvement. What would be the result if the context function was estimated in terms of information classes? We shall now address this question.

Table 10. Comparison of spectral and information class classification options using spectral class context, simulated data set 2a, reference classification as context template.

Classification	Information Class Class'n Accuracy, %		Spectral Class Class'n Accuracy, %	
	Overall	Ave.-by-Class	Overall	Ave.-by-Class
uniform-priors non-contextual	72.1	78.2	70.4	77.5
estimated-priors non-contextual	87.8	65.6	87.5	65.4
two-nearest-neighbors (north and east)	93.2	78.5	93.0	78.4
four-nearest-neighbors	97.1	87.5	97.1	87.5
eight-nearest-neighbors	98.2	92.0	98.2	92.0

Information-Class Context and Spectral-Class Classification

Up to this point we have assumed spectral-class context carries more usable contextual information than information-class context. It may be the case, though, that the information-class context carries most of the contextual information. Also, for the common case where the number of spectral

classes may be half or a third the number of spectral classes, estimating over information classes rather than spectral classes leads to a large reduction of dimensionality of the context function. The large dimensionality of the context function in the spectral class formulation may in and of itself be a significant source of estimation error due to our attempting to estimate the large number of elements in the context function from too small of a sample. If this is indeed the case, the lower dimensionality of the context function estimated over information classes should lead to a more accurate estimate. The combination of the higher accuracy attainable with the information-class context function estimate and the possibility that information classes carry most of the contextual information may lead to more accurate classifications when information-class context is used.

As before, let the set $\Omega=\{\omega_1,\omega_2,\ldots,\omega_m\}$ represent spectral classes and let the set $\Gamma=\{\gamma_1,\gamma_2,\ldots,\gamma_n\}$, $n \leq m$, represent information classes. Let $\underline{\vartheta}^p \in \Omega^p$ and $\underline{\xi}^p \in \Gamma^p$ stand for p-vectors of classes over spectral and information classes, respectively. If we assume that the spectral classes carry no contextual information outside of that carried by their information-class membership, we can calculate the context function over spectral classes, $G(\underline{\vartheta}^p)$, from the context function over information classes, $H(\underline{\xi}^p)$, as follows:

$$G(\underline{\vartheta}^p) = \sum_{\underline{\xi}^p \in \Gamma} H(\underline{\xi}^p) \prod_{k=1}^{p} p(\vartheta_k \mid \zeta_k). \tag{33}$$

The weights, $p(\vartheta_k \mid \zeta_k)$, represent the relative frequency of observing a spectral class, ϑ_k, given that a particular information class was observed. Inserting equation (33) into equation (9) gives the decision rule for information-class context and spectral-class classification (option 3), *viz:*

$d(\underline{X}_{ij})$ = the action $\alpha \in \Omega$ which maximizes

$$\sum_{\substack{\underline{\vartheta}^p \in \Omega^p, \\ \vartheta_p = \sigma}} \left\{ \sum_{\underline{\zeta}^p \in \Gamma} H(\underline{\zeta}^p) \prod_{k=1}^{p} p(\vartheta_k \mid \zeta_k) \right\} \prod_{k=1}^{p} f(X_k \mid \vartheta_k) \tag{34}$$

We might expect that spectral classes do carry some contextual information outside of their information-class membership. If this were the case we should observe that, if the context function estimates are very accurate, the spectral-class estimate would produce better results than the information-class estimate using equation (33) when used in the contextual decision rule (9). This is precisely what happens when the context functions are determined directly from the reference classification for the simulated data set 2a. Using two neighbor context (north and west neighbors), the spectral-class estimate produced overall and average-by-class accuracies of 93.0 and 78.4 percent. The corresponding information-class estimate result was 91.2 and 74.0 percent. As expected, the information-class estimate produced a significantly less accurate classification.

When a less accurate estimate of the context function is used, one might expect that the information-class estimate would produce more accurate classification results. This is what happened when the uniform-priors non-contextual classification was used to form the context function estimate for simulated data set 2a. Using two-neighbor context (north and west neighbors), the spectral-class estimate of the context function produced overall and average-by-class accuracies of 78.4 and 81.1 percent. The corresponding information-class estimate result was 79.8 and 81.7 percent.

These simulated data results show that the information-class estimate of the context function produces less accurate classifications than those

produced with a spectral-class estimate when the context function is known very accurately. However, the information-class estimate produces more accurate classifications when the context function must be estimated less accurately as from a uniform-priors non-contextual classification. This indicates that the information-class estimate is sufficiently less sensitive to errors from an imprecise estimate of the context function so as to produce better results despite any additional information spectral-class context may carry.

The first real-data test was performed using the Bloomington, Indiana, data set. For two-neighbor context (north and west neighbors), the spectral-class estimate produced overall and average-by-class accuracies of 84.5 and 84.2 percent. The corresponding information-class estimate result was 85.9 and 85.8 percent. These results are quite similar to the two-neighbor simulated data-results.

A test was also performed using four-nearest-neighbor context. The spectral-class context function calculated from the information-class estimate by equation (33) had to be thresholded in this case, i.e., context vectors, $\underline{v}^p$, with relative frequency of occurance less than a threshold value (here 6×10^{-5}) were eliminated from the sum in equation (34). If a nonthresholded context function were used here, there would be so many separate context vectors to sum over in equation (34) that the computer program would take an impractical amount of time, even over a small 50-pixel-square test area. The four-nearest-neighbor spectral class estimate produced overall and average-by-class accuracies of 84.5 and 84.1 percent. The information-class estimate produced accuracies of 88.2 and 88.7 percent

The same tests were repeated using the LACIE data set. For two-neighbor context (north and west neighbors), the spectral-class estimate produced overall and average-by-class accuracies of 80.0 and 72.1 percent. The corresponding information-class estimate produced accuracies of 80.4 and 72.4 percent. This accuracy improvement is much smaller than that obtained with the Bloomington, Indiana, data set, and may not even be statistically significant. In the four-nearest-neighbor-context case, two different information-class estimates (one thresholded at 6×10^{-5}, the other at 4×10^{-5}) produced lower accuracies than did the spectral-class estimate.

Before we attempt to draw any further conclusions from these results, we should investigate the remaining option in the spectral-class-versus-information-class question. This option (option 4) estimates the context function over information classes as does the option just discussed, but it also classifies into information classes rather than spectral classes.

Information-Class Context and Information-Class Classification

When the contextual classifier decision rule was derived in Chapter II, the set Ω and the p-vector $\underline{v}^p$ were not restricted to be spectral classes as they have been in this chapter. If Ω is replaced by Γ and $\underline{v}^p$ is replaced by $\underline{\xi}^p$, the desired information-class formulation of the decision rule follows directly from a derivation identical to that leading to equation (9):

$$d(\underline{X}_{ij}) = \text{the action } \alpha \in \Gamma \text{ which maximizes}$$

$$\sum_{\substack{\underline{\xi}^p \in \Gamma^p, \\ \xi_p = \alpha}} H(\underline{\xi}^p) \prod_{k=1}^{p} g(X_k \mid \xi_k). \tag{35}$$

Here $H(\underline{\xi}^p)$ is the context function over information classes, the $g(X_k \mid \xi_k)$ are

the information-class-conditional densities, and ζ_p is the p^{th} element of $\underline{\zeta}^p$. Under the usual methods of estimation, the density $g(X_k \mid \zeta_k)$ is a weighted sum of normal densities, *viz.*,

$$g(X_k \mid \zeta_k) = \sum_{\vartheta_k \in \zeta_k} p(\vartheta_k \mid \zeta_k)\, f(X_k \mid \vartheta_k) \tag{36}$$

where the $p(\vartheta_k \mid \zeta_k)$ are as in equation (33).

An information-class formulation of the contextual classifier decision rule identical to that given in equation (35) can be arrived at from a different perspective. The contextual classification decision rule defined by equation (32) classifies over information classes as does equation (35). The context function, $C(\underline{\vartheta}^p)$, used in equation (32) was assumed to be estimated directly from a spectral class template. If, rather, the spectral-class context function, $C(\underline{\vartheta}^p)$, is calculated from $H(\underline{\zeta}^p)$ using (33), equation (32) becomes:

$$d(\underline{X}_{ij}) = \text{the action } \alpha \in \Gamma \text{ which maximizes } d_\alpha(\underline{X}_{ij})$$

where

$$d_\alpha(\underline{X}_{ij}) = \sum_{\sigma \in \alpha} \left[\sum_{\substack{\underline{\vartheta}^p \in \Omega^p, \\ \vartheta_p = \sigma}} C(\underline{\vartheta}^p) \prod_{k=1}^{p} f(X_k \mid \vartheta_k) \right]$$

$$= \sum_{\sigma \in \alpha} \left[\sum_{\substack{\underline{\vartheta}^p \in \Omega^p, \\ \vartheta_p = \sigma}} \left\{ \sum_{\underline{\zeta}^p \in \Gamma} H(\underline{\zeta}^p) \prod_{k=1}^{p} p(\vartheta_k \mid \zeta_k) \right\} \prod_{k=1}^{p} f(X_k \mid \vartheta_k) \right]$$

$$= \sum_{\substack{\underline{\zeta}^p \in \Gamma^p, \\ \zeta_p = \alpha}} H(\underline{\zeta}^p) \left[\sum_{\underline{\vartheta}^p \in \underline{\zeta}^p} \prod_{k=1}^{p} p(\vartheta_k \mid \zeta_k) f(X_k \mid \vartheta_k) \right]$$

$$= \sum_{\substack{\underline{\zeta}^p \in \Gamma^p, \\ \zeta_p = a}} H(\underline{\zeta}^p) \left[\prod_{k=1}^{p} \sum_{\vartheta^p \in \underline{\zeta}^p} p(\vartheta_k \mid \zeta_k) f(X_k \mid \vartheta_k) \right]$$

$$= \sum_{\substack{\underline{\zeta}^p \in \Gamma^p, \\ \zeta_p = a}} H(\underline{\zeta}^p) \prod_{k=1}^{p} g(X_k \mid \zeta_k).$$

which is identical to equation (33) as suggested. It proved initially to be more convenient to implement the decision rule given in equation (35) by implementing equation (32) and calculating $G(\underline{\vartheta}^p)$ using equation (33). This was because the program implementing the original pure spectral-class formulation could be trivially modified to implement equation (32), and a small program written to calculate the spectral-class context function from the information-class context function using equation (33).

The classification results obtained using the information-class formulation (option 4) are compared in Tables 11 and 12 with those obtained using other formulations. In Tables 11 and 12, options 3 and 4 show nearly identical results. This is consistent with the results shown in Table 10 where options 1 and 2 gave nearly identical results. (Option 2 was not tested in Tables 11 and 12 for this reason.) These results show that information-class classification produced nearly identical results as those produced by the spectral-class classification irregardless of whether information-class or spectral-class context was employed.

Tables 11 and 12 also show that information-class context generally produced better classification results. This result is consistent with the expectation expressed in the discussion above about the relative merits of information-class and spectral-class context. For an inaccurate method of context function estimation such as the classify-and-count method, we

Table 11. Comparison of spectral- and information-class classification and context options, Bloomington, Indiana, data set, uniform-priors non-contextual classification as context template.

Context	Option	Accuracy, %	
		Overall	Ave.-by-Class
uniform-priors non-contextual	(-) spectral-class class'n	83.1	82.7
two-nearest-neighbors (north and west)	(1) spectral-class context and spectral-class class'n	84.5	84.2
"	(3) information-class context and spectral-class class'n	85.9	85.9
"	(4) information-class context and information-class class'n	85.7	85.8
four-nearest-neighbors	(1) spectral-class context and spectral-class class'n	84.5	84.1
"	(3) information-class context and spectral-class class'n	88.2	88.7
"	(4) information-class context and information-class class'n	87.9	88.2

expected that information-class context would produce better classification results.

Earlier we noted that information-class context produced better classification results with the unbiased estimation method, while spectral-class context produced better results with the ground-truth-guided method. This result is consistent with the discussion and results of this chapter. Since for the tests performed on the ground-truth-guided method and the unbiased estimation method, the ground-truth-guided method produced the best classification results, we would expect that the spectral-class formulation would perform relatively better for the ground-truth-guided method than for the unbiased estimation method.

Table 12. Comparison of spectral- and information-class classification and context options, LACIE data set, uniform-priors non-contextual classification as context template.

Context	Option	Accuracy, %	
		Overall	Ave.-by-Class
uniform-priors non-contextual	(-) spectral-class class'n	78.7	72.0
two-nearest-neighbors (north and west)	(1) spectral-class context and spectral-class class'n	80.0	72.1
"	(3) information-class context and spectral-class class'n	80.4	72.4
"	(4) information-class context and information-class class'n	80.6	72.6
four-nearest-neighbors	(1) spectral-class context and spectral-class class'n	79.6	72.1
"	(3) information-class context and spectral-class class'n	78.3	71.5
"	(4) information-class context and information-class class'n	78.2	71.4

CHAPTER VII - PREDICTING THE OPTIMAL P-CONTEXT ARRAY

Prior to the development of the unbiased estimator, methods were sought with which to improve the practical effectiveness of the classify-and-count and power methods for estimating the context function. For both of these methods, it was noticed that a smaller p-context array (p = 2 or 3) was generally more effective in early iterations. For general scenes, nearest-neighbors seem to provide the most useful contextual information, but when context arrays of fewer than four nearest neighbors are used, it is not clear which neighbors should be used. The practical effectiveness of the classify-and-count and power methods could be improved if an effective predictor of the optimal p-context array could be found.

One could discover the optimal p-context arrays at each iteration by simply performing a large number of contextual classifications over a training set. This could be quite time consuming, however. A more desirable solution would be to predict the optimal p-context array at each iteration from some characteristic of the data such as a "context measure" before actual classifications are performed.

Suppose that the context function, $C(\underline{v}^p)$ is such that it can be written in product form, i.e.,

$$C(\underline{v}^p) = C_1(\underline{v}')\cdot C_2(\underline{v}'') \tag{37}$$

where $\underline{v}'$ and $\underline{v}''$ are, respectively, q and p-q vectors of classes. The elements

of $\underline{\vartheta}'$ are identical to the first q elements of $\underline{\vartheta}^p$, and the elements of $\underline{\vartheta}''$ are identical to the last p-q elements of $\underline{\vartheta}^p$. If this factorization can indeed be realized, equation (9) can be rewritten as

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

$$\left[\sum_{\underline{\vartheta}' \in \Omega^q} C_1(\underline{\vartheta}') \prod_{k=1}^{q} f(X_k \mid \vartheta_k)\right] \cdot \left[\sum_{\substack{\underline{\vartheta}'' \in \Omega^{p-q} \\ \vartheta_p = a}} C_2(\underline{\vartheta}'') \prod_{k=q+1}^{p} f(X_k \mid \vartheta_k)\right] \tag{38}$$

where the ϑ_k, k=1,2,...,p, are the elements of $\underline{\vartheta}^p$. Since the term in the first set of brackets is independent of the decision a, it is just a constant factor relative to the decision process and can be ignored when classifying the point at (i,j).

If $C(\underline{\vartheta}^p)$ can be factored as in equation (37), then $\underline{\vartheta}'$ and $\underline{\vartheta}''$ are statistically independent. This suggests that a measure of departure from independence of $\underline{\vartheta}'$ and $\underline{\vartheta}''$ may be useful as a measure of additional contextual information carried by the pixel positions in $\underline{\vartheta}'$ over that carried by the pixel positions in $\underline{\vartheta}''$. One measure of this departure is

$$\Delta C_q^p \triangleq \sum_{\underline{\vartheta}^p \in \Omega^p} \left[C_1(\underline{\vartheta}') \cdot C_2(\underline{\vartheta}'') - C(\underline{\vartheta}^p)\right]^2 \tag{39}$$

where $C_1(\underline{\vartheta}')$ and $C_2(\underline{\vartheta}'')$ are marginals of $C(\underline{\vartheta}^p)$. Thus the departure of the factorization of $C(\underline{\vartheta}^p)$ into its marginals from a true factorization is here defined as the "context measure" ΔC_q^p.

To investigate the use of the context measure ΔC_q^p in predicting the optimal p-context array we use the following approach. Establish $\underline{\vartheta}''$ as a fixed (p-q)-dimensional classification vector which we shall call the "core

array". Calculate the values of ΔC_q^p for various q-dimensional classification vectors $\underline{v}'$ with elements distinct from the core array. Only those q-dimensional arrays that are expected to add significant contextual information need be investigated. The best p-context array would be the (p-q) pixel locations of $\underline{v}''$ combined with the q pixel locations of the $\underline{v}'$ that produced the largest value for ΔC_q^p. Of course, this assumes that the contextual information contributed by the $\underline{v}'$ pixel locations is not so erroneous as to actually decrease classification accuracy. This may not be a reasonable assumption in all cases as we will see in some of the real data tests that are reported later in this chapter.

ΔC_q^p was tested as a context measure to predict the best p-context array in terms of relative pixel locations as shown in Figure 14. Usually pixel location 5 was the pixel to be classified. In some cases pixel location 1 was used as the pixel to be classified.

1	2	3
4	5	6
7	8	9

Figure 14. Pixel locations used in testing ΔC_q^p.

The first test of ΔC_q^p was performed on the simulated data with spectral-class context functions estimated by tabulation from a the reference classification (the "ground truth"). One-neighbor context was considered. As can be seen in Table 13, ΔC_q^p clearly predicted that the best neighbor to use

for context would be one of the four nearest neighbors (pixel positions 2, 4, 6 or 8). It was not conclusive from the tabulated results whether any particular nearest neighbor was better than the others as context. Nevertheless, this test seemed to indicate that ΔC_q^p works quite well when the context is perfectly known.

Table 13. ΔC_q^p tested on simulated data with context functions determined from reference classification.

$\underline{\vartheta}'$	$\underline{\vartheta}''$		Accuracy, %	
Pixel Location	Pixel Location	$\Delta C_1^2 \times 10^4$	Overall	Average-by-Class
8	5	5.09	92.7	74.0
2	5	4.99	91.6	73.5
4	5	4.90	91.7	71.8
6	5	4.90	91.7	73.9
7	5	3.42	90.8	71.2
3	5	3.31	90.4	69.8
9	5	3.26	90.6	70.6
1	5	3.19	90.6	70.1
7	1	2.58	90.3	68.6
3	1	2.27	90.2	70.3
8	1	1.98	89.4	67.9
6	1	1.87	90.4	70.2
9	1	1.53	89.9	69.5

ΔG_q^p was tested again on simulated data, but with the context function estimated using the classify-and-count method. Here the context should still be fairly accurate, since the classify-and-count method did perform well on the simulated data set. Table 14 shows that ΔG_q^p correlates fairly well with classification accuracy.

Table 14. ΔG_q^p tested on simulated data with context functions estimated from uniform-priors non-contextual classification.

$\underline{v}'$	$\underline{v}''$		Accuracy, %	
Pixel Location	Pixel Location	$\Delta G_1^2 \times 10^6$	Overall	Average-by-Class
8	5	7.56	79.8	81.7
2	5	7.30	79.1	81.9
4	5	6.13	78.8	80.6
6	5	6.11	79.0	81.4
7	5	4.71	78.8	80.9
3	5	4.53	78.6	80.6
9	5	4.28	78.4	80.6
1	5	4.22	78.3	79.7
7	1	3.77	78.5	80.9
8	1	2.73	78.0	80.0
3	1	2.65	78.0	80.9
6	1	2.31	78.0	80.8
9	1	2.17	78.0	80.1

The first real-data test of ΔC_η^p was performed on the Bloomington, Indiana, data set described in Chapter III. The results are displayed in Table 15. Here the differences in the value of the context measure ΔC_1^2 were not well correlated with the accuracy of the classification results. Similar results were seen in a test using the LACIE data set described in Chapter IV. It may be that in these real data cases, the context as estimated from the non-contextual classification is not sufficiently accurate for the context measure to function properly as a predictor of the best p-context array.

Table 15. ΔC_η^p tested on Bloomington, Indiana, Landsat data set. Context functions estimated from uniform-priors non-contextual classification.

$\underline{v}'$	$\underline{v}''$		Accuracy, %	
Pixel Location	Pixel Location	$\Delta C_1^2 \times 10^5$	Overall	Average-by-Class
4	5	7.69	84.2	83.8
6	5	7.68	84.8	84.1
2	5	5.40	85.2	84.8
8	5	5.31	83.8	83.4
3	5	3.79	84.2	83.8
7	5	3.61	84.0	83.5
1	5	3.04	84.4	84.1
9	5	2.96	83.7	83.2

Tests with the power method were performed on the two real data sets to see how significant this failure of ΔC_q^p to predict some best p-context array is in these cases. Table 16 summarizes the results of two iterations of the power bootstrap method in which various two-neighbor contexts were used in the first iteration. Four-nearest-neighbor context was used for the second iteration.

Table 16. Power method results for various pixel locations of the two-neighbors used for first iteration context. Classified pixel location is location 5. Second iteration uses four-nearest-neighbor context.

Data Set	1st Iteration Context Pixel Locations	Best Power 1st Iteration	Best Power 2nd Iteration	2nd Iteration Accuracy, % Overall	2nd Iteration Accuracy, % Average-By-Class
LACIE	2 & 4	15	10	86.7	75.6
LACIE	2 & 8	15	10	86.7	75.6
LACIE	4 & 6	15	10	86.7	75.6
Bloomington	2 & 6	10	5	88.5	87.5
Bloomington	2 & 8	10	5	88.6	87.8
Bloomington	4 & 6	7	3	88.2	88.2
Bloomington	4 & 8	10	5	89.7	89.2
Bloomington	3 & 7	7	3	87.2	87.1

For nearest-neighbor context, the choice of 1st iteration context makes virtually no difference for the LACIE data set in terms of 2nd iteration accuracies. There are some differences in the Bloomington data set results. As might be expected, the non-nearest-neighbor case (1st iteration pixel locations 3 and 7) produced a lower 2nd iteration accuracy. It would not be expected from the results of Table 15 that nearest-neighbor pixel locations 4

and 8 would produce better classification accuracies.

It should be remembered that the Bloomington data set results are evaluated from just over half the pixels in the 50-pixel square scene (1317 pixels) while the LACIE data set is evaluated from ground truth over the entire 50-pixel square scene. Also, the Bloomington data set ground truth was derived from aircraft infrared photography while the LACIE ground truth was from a ground survey. The combination of these facts may serve to make the Bloomington data set results sufficiently noisy to make the variations in the accuracies displayed in Table 16 are not statistically significant.

If indeed no one particular nearest neighbor is better as context in these two real data cases, it remains to be explained why ΔC_q^p produced a larger value for pixel locations 4 and 6 versus pixel locations 2 and 8 on the Bloomington data set (Table 15) and on the LACIE data set (not shown). An interesting fact that comes to mind is that the Landsat sampling rate is significantly finer in the across-track direction than for the along-track direction. The neighboring pixels which are geographically closer to the pixel in question should show more statistical correlation to that pixel than those neighbors at a larger geographical distance. Thus, we should expect that ΔC_q^p would produce larger values for the pixels in the across-track direction (pixel locations 4 and 6) than for the pixels in the along-track direction (pixel locations 2 and 8) from Landsat sampling characteristics alone. Unfortunately, the sampling difference reflected in the values of ΔC_q^p had no consistent effect on the performance of individual nearest-neighbor pixels as context for contextual classification.

The above results indicate that ΔC_q^p is not a useful predictor of the optimal p-context array. However, the results presented in Table 11 suggest

that such a predictor may not even be necessary for the optimal use of the classify-and-count and power methods. Also, in Chapter IV we saw that the ground-truth-guided and unbiased context function estimation methods performed consistently well with four-nearest-neighbor context. All of these results tend to obviate the need for a predictor of the optimal p-context array.

CHAPTER VIII - SUMMARY AND DIRECTIONS FOR FURTHER RESEARCH

This paper has explored the theoretical basis and implementation of a general statistical classification decision rule which exploits both spatial and spectral information when classifying multispectral image data. A contextual classifier based on this decision rule depends only on general contextual information, and can, in principle, be used to advantage on any remotely-sensed multispectral image data set.

Summary of Results

The theoretical derivation of the contextual decision rule was presented in Chapter II. This theoretical development was an elaboration and clarification of a development given by Swain and Vardeman in [3]. It was noted in Chapter II that the optimal decision rule cannot be implemented in practice since it depends on the context function, $G(\underline{\vartheta}^p)$, and the class-conditional densities, $f(X_k \mid \vartheta_k)$, which are unknown. Thus, the performance of the contextual classifier depends directly on how well $G(\underline{\vartheta}^p)$ and the $f(X_k \mid \vartheta_k)$ can be estimated.

Methods for estimating the class-conditional densities are well established from considerable experience with the non-contextual maximum likelihood decision rule. One of the principal research topics of this paper has been the development of effective and practical methods for estimating the context function. A simple method for estimating the context function, the classify-and-count method, was explored in Chapter III in tests on simulated

and real Landsat data sets. The results of these early exploratory experiments pointed to the three main areas of research described in the remaining chapters of the paper.

The poor performance of the classify-and-count method on real Landsat data sets pointed to the need for a better context function estimation method. Speculation on the reasons for the inadequacy of the classify-and-count method led to the formulation of two alternative methods: the ground-truth-guided method and the power method (Chapter IV). The reported tests have shown the ground-truth-guided method to be an effective and practical method, provided that sufficient ground truth is available in spatially contiguous blocks. While the power method does not need such special ground truth and can provide significant improvements in classification accuracy, the power method turned out to be impractical to use. An unsuccessful attempt to develop a context measure to use in conjunction with the power method (and the classify-and-count method) to improve its practicality was described in Chapter VI.

For cases where sufficient spatially contiguous ground truth is not available for estimating the context function, an unbiased estimation method was developed (Chapter IV). This unbiased estimator has the additional advantage of being amenable to an adaptive implementation, so that the resulting context function estimate is more closely tailored to local conditions in the image data.

The second research problem area suggested by the early experimental results is the need to reduce the computational complexity of the contextual classifier. An approximate algorithm was developed (Chapter V) which requires less than half of the computer time taken by the original implementation in the tests performed. A faster hybrid algorithm was also suggested in

Chapter V but is not yet perfected. It was further noted in Chapter IV that a pure information-class formulation of the contextual classifier is significantly less computationally intensive than a formulation involving spectral classes.

The third research problem area involved certain assumptions made in the original implementation of the contextual classifier. Chapter VI explored in detail the relative merits of using spectral classes or information classes as the basis of context function estimation and classification when using the classify-and-count method. The conclusion drawn was that in this case, for real Landsat data sets, the contextual classifier performed better when the context function was estimated in terms of information classes. No significant difference in performance was observed when the classification was done in terms of spectral classes or in terms of information classes. In Chapter IV we noted that a pure spectral-class formulation performed slightly better with the ground-truth-guided method and that a pure information-class formulation performed best with the unbiased estimator. This question will be mentioned again in the discussion of directions for further research.

A second assumption included in the third research area was the class-conditional independence assumption represented by equation (8) in Chapter II. This assumption has yet to be studied (see below).

Directions for Further Research

The research presented in this paper suggests further study in two directions. One would be to pursue the theoretical foundations of the contextual classifier, in particular the effect of the class conditional independence assumption. Another direction of study would be to investigate a practical implementation of the contextual classifier which can be used effectively with data sets larger than the 50-pixel-square data sets employed throughout the

present study. We address the implementation question first.

Two particular implementations of the contextual classifier are good candidates for further study. These are implementations which use (a) the ground-truth-guided method and (b) an adaptive version of the unbiased estimation method to estimate the context function. In either case, the approximate algorithm should be employed. Research into the hybrid algorithm should be pursued and, if research results are favorable, this algorithm should be incorporated into the implementation.

Implementation Using the Ground-Truth-Guided Method. On the two 50-pixel-square data sets tested, the ground-truth-guided method produced classification accuracies significantly better than those produced using the unbiased estimation method. It should be noted, however, that in these two cases fully one-half of the data set was designated as the training set for the ground-truth-guided method. In practical classification problems using much larger data sets, it is usually the case that ground truth is available for only ten percent or less (often less than one percent) of the data set. We expect that this smaller percentage of ground truth data will decrease the effectiveness of the ground-truth-guided method.

As noted earlier, the spectral-class formulation of the ground-truth-guided method produced somewhat higher classification accuracies than the information-class formulation. Because the information-class formulation requires less than half the computer time required by the spectral-class formulation, this becomes a factor of importance for larger data sets. If the information-class formulation continues to give poorer classification results for larger data sets, it should be attempted to discover a variation on the present information-class formulation that does not give poorer results. However, we expect that on larger data sets the present information-class

formulation will produce higher classification accuracies than those produced by the spectral-class formulation. As noted in the previous paragraph, the ground-truth-guided method may not produce as accurate an estimate of the context function for larger data sets. This is likely to cause the information-class formulation to perform relatively better as it is less sensitive to estimation errors (see Chapter VI).

Implementation Using the Unbiased Estimator. The present adaptive information-class formulation of the unbiased estimator requires significantly less computer time than the other formulations tested. This is because this formulation produces fewer non-zero elements in the estimate of the context function than is the case for any other formulation. Further, the adaptive information-class formulation gave either approximately the same or significantly better classification accuracies than any other unbiased-estimator formulation. One question that needs to be resolved for the adaptive information-class formulation for a larger, practical-sized data set is the selection of generally optimal classification and estimation data block sizes. For the three small-scale data sets tested, estimating the context function from a 20, 25, or 35-pixel-square block of data centered on the corresponding 10, 17, or 25-pixel-square classification block seemed to be optimal depending on the data set tested. It remains to be seen whether one particular choice of data block size will be nearly optimal for most or all larger data sets. Fortunately, classification accuracies do not seem to be highly sensitive to the size of the data blocks chosen.

Although the present version of the adaptive information-class formulation uses less computer time than other formulations of the unbiased esimator, the present version can still be improved substantially in this regard by removing redundant calculations and storing the context function estimates

in main memory rather than writing the estimated relative frequencies in an external file. It should be noted that, for even moderate values of p (the number of pixels in the p-context array), storing the context function estimate in main memory would be impossible if a spectral class formulation were used. There would not be enough space to store all the non-zero entries of a spectral-class context function.

The Class-Conditional Independence Assumption. The original derivation of the contextual classification algorithm assumed class-conditional independence among all image locations. It would be of interest to investigate the implications of this assumption. A method for experimentally investigating these implications is outlined below.

For contextual classifications using an arbitrary p-context array, the class-conditional density $f(\underline{X}_{ij} \mid \underline{v}^p)$ of equation (7) could be estimated by clustering in a manner similar to the way the densities $f(X_k \mid v_k)$ of equation (9) are estimated (see [1]). In this case, however, the clustering would be done based on the $n \times p$ dimensional $\underline{X}_{ij}$ rather than the n-dimensional X_k. Significant clusters of the observation vectors, $\underline{X}_{ij}$, could then be identified with a particular classification vector, $\underline{v}^p$, and the multivariate normal approximation for $f(\underline{X}_{ij} \mid \underline{v}^p)$ could be used. Clustering done in such a way would provide class-conditional densities $f(\underline{X}_{ij} \mid \underline{v}^p)$ without an independence assumption for use in comparison to classifier tests using class-conditional densities assumed to be independent among all image locations.

The use of the class-conditional density $f(\underline{X}_{ij} \mid \underline{v}^p)$ presents the practical problem of effectively working with a multispectral data set with a very large number of channels. Some of the dimensionality reduction techniques used in working with other large-dimensioned data sets may be necessary in this case.

LIST OF REFERENCES

[1] P. H. Swain and S. M. Davis, eds., Remote Sensing: The Quantitative Approach, McGraw-Hill, New York, 1978.

[2] D. A. Landgrebe, "The Development of a Spectral-Spatial Classifier for Earth Observation Data," Pattern Recognition, Vol. 12, No. 3, pp. 165-175, May-June 1980.

[3] P. H. Swain, S. B. Vardeman and J. C. Tilton, "Contextual Classification of Multispectral Image Data," LARS Technical Report 011080, Laboratory for Applications of Remote Sensing (LARS), Purdue University, West Lafayette, Indiana 47907, January 1980.

[4] 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.

[5] H. Robbins, "Asymptotically Subminimax Solutions of Compound Statistical Decision Problems," Proc. Second Berkeley Symp. Mathematical Statistics and Probability, pp. 131-148, University of California Press, 1951.

[6] J. Van Ryzin, "The Compound Decision Problem With m × n Finite Loss Matrix," Annals of Mathematical Statistics, Vol. 37, pp. 412-424, 1966.

[7] J. Van Ryzin, "The Sequential Compound Decision Problem With m × n Finite Loss Matrix," Annals of Mathematical Statistics, Vol. 37, pp. 954-975, 1966.

[8] S. Vardeman, "Admissible Solutions of k-Extended Finite State Set and Sequence Compound Decision Prolems," Journal of Multivariate Analysis, Vol. 10, pp. 426-441, September 1980.

[9] 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 (LARS), Purdue University, West Lafayette, Indiana 47907, May 1975.

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

[11] 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 × n Components," Research Memo RM-333, Statistics and Probability, Michigan State University, 1975.

[12] E. F. Kit and P. H. Swain, "An Approach to the Use of Statistical Context in Remote Sensing Data Analysis," Proceedings of the Fifth Canadian Symposium on Remote Sensing, Victoria, B. C., August 1978.

[13] J. Hannan, D. Gilliland and S. B. Vardeman, "Empirical Bayes and Compound Decision Theory: A Survey and Annotated Bibliography," (in preparation).

[14] B. W. Smith, H. J. Siegel and P. H. Swain, "Contextual Classification on a CDC Flexible Processor System," Machine Processing of Remotely Sensed Data (IEEE Catalog No. 1981 CH 1637-8 MPRSD), September 1981.

[15] M. Alvo and M. Goldberg, "A Measure of Reliability for Classification of Earth Satellite Data," IEEE Trans. Systems, Man and Cybernetics, Vol. SMC-11, No.4, pp. 312-318, April 1981.

[16] M. M. Hixson, D. K. Scholz, N. C. Fuhs and T. Akiyama, "Evaluation of Several Schemes for Classification of Remotely Sensed Data," Photogrammetric Engineering and Remote Sensing, Vol.46, No. 12, pp. 1547-1553, December 1980.