Reprinted from

Eighth International Symposium

Machine Processing of

Remotely Sensed Data

with special emphasis on

Crop Inventory and Monitoring

July 7-9, 1982

Proceedings

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

EVALUATION OF THE APPLICATION OF LANDSAT DATA TO CROP DISCRIMINATION IN WESTERN AUSTRALIA

N.A. CAMPBELL, F.R. HONEY, P.T. HICK, M.D.W. CARLTON

Commonwealth Scientific and Industrial Research Organization/Divisions of Mathematics and Land Resources Management

I. ABSTRACT

This paper describes progress results from an ongoing collaborative research project between Cooperative Bulk Handling (CBH) and the Commonwealth Scientific and Industrial Research Organization (CSIRO) to evaluate the role of remotely sensed data in crop production forecasting in Western Australia.

Seventeen areas have been monitored throughout the wheatbelt of Western Australia (WA). Computer-compatible tapes for all cloud-free satellite overpasses from 18 June 1981 to 29 November 1981 have been obtained.

The project has three distinct phases:

i) discrimination of grain crops from pasture, and determination of the area planted to crop; ii) discrimination between crop types; and iii) estimation of yield.

To date, most effort has been applied to estimation of the area planted to crop, and to the discrimination of grain crops from pasture. Complications which have arisen due to variations in planting practice and the effect of this variation on the identification of areas planted to crop are discussed in the paper. The application of temporal analysis and growth curve analysis to improve the discrimination of crops from pasture and hence the estimation of area planted to crops, and the discrimination between crop types, is described.

II. INTRODUCTION

The Western Australian wheatbelt, situated in the south-west corner of Australia (Figure 1), covers an area slightly in excess of 6 million hectares, with a yearly rainfall range between 300 and 600 mm. Wheat planted after early winter rains accounts for approximately 85% of the area planted to crop. The harvest, gathered between October and December, varies between 3.5 and 6 million tonnes. Early estimates of area planted to crop, of crop type and of potential yield are necessary to enable planning for regional and central storage, transport and marketing. Crop inventory and yield estimation techniques which utilize satellite data could complement existing survey and estimation methods, and improve their accuracy.

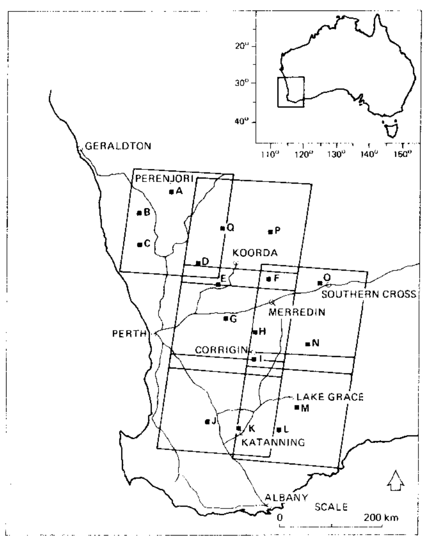

Figure 1. Location of Wheatbelt and Distribution of Study Sites. Approximate coverage of six Landsat scenes os shown.

Projects such as LACIE (Large Area Crop Inventory Experiment) in the United

States which use satellite data from different times in the growing season to monitor crops have produced impressive results. Encouraging results have also been reported by the New South Wales Department of Agriculture in a collaborative project with IBM Australia[6].

A problem in applying techniques such as those developed during LACIE, and being developed in AgRISTARS, is the variability which occurs in the wheatbelt of Western Australia. Large variations in soil type occur across the whole wheatbelt, and may even occur within paddocks. The rainfall distribution is non-uniformly distributed in time and intensity. Farm management practices such as planting time and planting technique, and fertilizer levels applied, differ between farms and between districts.

A pilot project was undertaken by CSIRO in 1978 to assess the role of Landsat data for crop discrimination in the wheatbelt of Western Australia. Only two cloud-free images were acquired for the areas studied. Results were promising for the identification of wheat and barley, though differences between and within paddocks of the same crop were sometimes as great as differences between paddocks of different crops.

A limitation of this early work was the lack of temporal data; the Australian Landsat Station did not commence operation until 1980. From early 1980, data from the Landsat satellites have been recorded for the whole of Australia every eighteen days. Also, these data can be acquired by a user within ten days of an overpass, potentially making the use of Landsat data a practical and timely means for assessing crops.

III. THE CURRENT PROJECT

The current project, which commenced in 1981, involves collaboration between the Remote Sensing Group in the Division of Land Resources Management, CSIRO, the Perth group of the Division of Mathematics and Statistics, CSIRO, and Co-operative Bulk Handling (CBH) and is designed to evaluate the capability of Landsat data to determine areas planted to crops, to discriminate between crops, and ultimately to estimate yield in the wheatbelt of Western Australia. [CBH are responsible for the collection, storage and transport of the wheat harvest in Western Australia.] Factors to be considered include variation in soil type, the spatial and temporal pattern of rainfall, and the individual management practices of farmers.

The project is designed to use the Landsat data for several overpasses during each growing season, register the scenes to the Australian map grid, and integrate the Landsat data with information gathered on soil, climatic and management variables. It was anticipated that a temporal analysis of the spectral signatures derived from sequential Landsat overpasses would result in acceptable identification accuracy.

The proposed project is divided into three distinct phases:

i) the discrimination of grain crops from pasture and the determination of the area planted to crop;

ii) discrimination between grain crops of different types; and

iii) estimation of projected yield.

This paper reports some preliminary results related to the first objective.

A map illustrating the areas chosen for the study is presented in Figure 1. Table 1 summarizes the distribution of farms chosen as training sites throughout the area. Seventeen locations have been included in the study. Sites within each location are comprised of 3-5 farms distributed throughout the locality. The sites were chosen by CBH district superintendents to provide a representative sample of soil types and management practices for each area.

Table 1. CSIRO/CBH Crop Forecasting Study - Training Sites.

Area	Sites	Actual Crop Paddocks
A	20	30
B	14	18
C	20	20
D	12	22
E	15	44
F	15	19
G	27	40
H	14	20
I	19	20
J	9	9
K	18	20
L	11	11
M	10	21
N	14	16
O	11	20
P	7	17
Q	8	13
Total	244	360

The seventeen locations are distributed over an area covered by six Landsat scenes (Figure 1). During 1981, twenty-six Landsat tapes were purchased for the period 18 June to 11 November. Cloud-free images are available for seven overpasses for site H, six overpasses for site I, and five overpasses for sites F and G.

IV. GROUND-TRUTH DATA

Detailed ground-truth data were collected by CBH district superintendents in consultation with the farmers involved in the study. To standardize and collate the ground-truth data, three sets of cards have been designed to provide detailed records of each paddock on each farm in the study. A master record card contains details of the site, including the name of the farmer, the Landsat scene, local government area and the nearest rail siding for delivery of grain. The availability of historic records of rainfall, fertilizer and cropping details are also recorded, together with details of the area planted or quantity of grain to be retained for stockfeed on the property.

Crop phenology cards record detailed information on crop development for each paddock in the study. Two sets of cards are kept - one by the farmer and one by the CBH district staff. Each card records the name of the farmer and the paddock name and number, and the grain crop, including variety and number of hectares planted. Each card then gives a pictorial record of the development of the crop: for wheat, the stages were: planting; emergence; 3-6 leaf stage; 6-8 leaf stage; fully tillered; jointing; flowering; and maturity. The farmer is asked to enter the date at which the illustrated growth stage is reached, together with details of weed growth (type, density, patchy or uniform cover, etc) and crop health (including insect damage, fungus or other disease). The CBH district staff visit each farm in the study fortnightly, and record the growth stage, weed growth and crop health at the time of the visit.

Soil type and paddock history cards record paddock details. Each card records the name of the farmer; paddock name, number and description; area in hectares; landscape position; terrain and microrelief; elevation; and aspect. The soils within each paddock were grouped according to their principal profile form[9], a system based upon depth of soil horizons, colour, texture, gravels, mottling, consistence and pH. The soils were mapped from the Atlas of Australian Soils[10], from the pre-release soils and landform mapping prepared by the Western Australian Department of Lands and Surveys, from air photo interpretation and from ground inspection. The date of clearing for each paddock and the paddock history for the period 1970-1981 are recorded. The paddock history includes use (crop type, pasture, fallow); fertilizer applied; yield; rainfall; and specific comments.

With such detailed ground data, it should be possible to identify the major sources of variation within and between paddocks of the same crop and of different crops.

V. DATA ANALYSIS

Two stages of analysis are considered in this paper. The first stage provides an overall description of the separation between the training paddocks, and in particular determines whether differences between paddocks sown to crop and paddocks sown to pasture are greater than differences between paddocks sown to the same crop. To maintain consistency, the term group will henceforth be used to refer to the data for a training paddock, while the term variables will be used to refer to the Landsat bands. The approach adopted here is to provide an ordination of the group means based on the first few canonical variates. Since the canonical vectors are chosen to maximize the variation between the training or reference groups relative to the variation within groups over all possible linear combinations of the variables, the description provided by the first few canonical variates will usually contain most of the relevant information. The associated canonical roots - the ratios of the between-groups sum of squares to the within-groups sum of squares for the resulting canonical variate scores - indicate the degree of group separation. Subsets of the variables may also be examined; the canonical roots and associated test statistics, such as the largest canonical root, Wilk's likelihood ratio, and Hotelling's trace statistic, are then compared with those based on all variables to determine whether most of the important information has been retained.

The ordination based on the resulting adequate subset of variables may be used to determine the effective number of training or reference groups required.

Once the reference groups have been established, the second stage involves allocation of the data from each of the reference groups, and allocation of data from additional groups if available. Because data from additional groups were not available at the time of writing, the approach adopted here uses leave-one-out calculations for the reference groups. Posterior probabilities are estimated using Bayes Theorem. There are compelling arguments in the statistical literature to suggest that the multivariate densities should be estimated by multivariate Student densities rather than by multivariate Gaussian densities[1,2,3,8]. There is also the question of whether equal or unequal covariance matrices should be assumed for the calculations. An index of typicality is also calculated for each pixel (see [2] for a formal definition). This index reduces to the probability associated with the individual squared Mahalanobis distance for the pixel. For leave-one-out calculations or for pixels from additional testing sites, the probability is calculated by referring the squared Mahalanobis distance to the F distribution.

To evaluate the effect of atypical pixels on the separation of the groups and on the allocation results, a robust canonical variate analysis is used to provide an ordination which is insensitive to atypical pixels[4,5].

There is in principle no reason why the analysis should not proceed directly to the second stage. However if this is done, the graphical summary of the relationships between crop and pasture types and between paddocks of the same crop or pasture type is not available. Moreover, while selection of an adequate subset(s) of variables for allocation based on a criterion of group separation is not necessarily optimum[7], it does provide a reasonable basis on which to proceed, and appears to give satisfactory results in practice.

VI. RESULTS

Since seven cloud-free overpasses are available for site H, this paper will concentrate on the analysis of data from that site. Analyses are presented for two sets of data from site H. The first set consists of paddocks for which detailed ground-truth data are available. In view of the promising results obtained for the separation of crops from pasture (see below), additional data were extracted for what appeared to be homogeneous areas on the usual three-band display of the data. These paddocks were tentatively labelled as crop or pasture on the basis of their visual similarity with ground-truthed areas which had been positively identified as crop or pasture.

A. RESULTS FOR GROUND-TRUTHED PADDOCKS

A plot of the group means for the first two canonical variates shows clear separation of the two pasture paddocks from the paddocks sown to crop (Figure 2). There is also some separation of the barley paddock, though the oats paddock appears to show little separation from the wheat paddocks. There is also some grouping of wheat paddocks into those sown early in the season (W4, W5, W6 on June 3) and those sown some 10 days later (W1, W2, W3). All wheat paddocks included in this part of the analysis were spray-seeded.

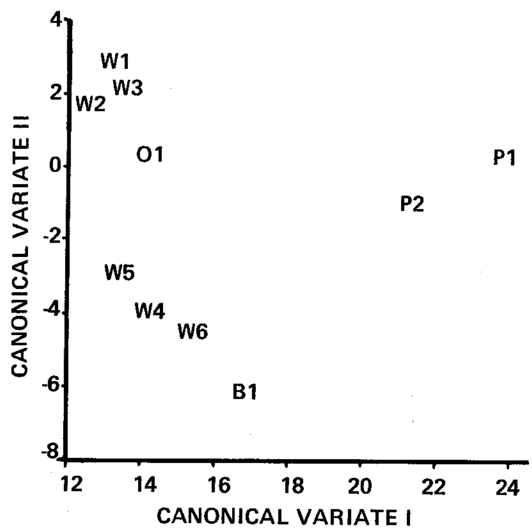

Figure 2. Paddock Means for First Two Canonical Variates For All Bands For All Dates. The canonical vectors are scaled to unit standard deviation within groups.

The first two canonical roots are 21.81 and 7.11, accounting for 91% of the overall between-groups variation. The relationships between the groups are little changed when either band 4 and band 6 or band 4 and band 7 for each of the seven overpasses are eliminated from the analysis (first two canonical roots of 18.47 and 6.59 versus 19.18 and 5.79 respectively); analyses based on either band 5 and band 7 or on band 5 and band 6 give similar ordinations, with that based on bands 5 and 7 showing marginally greater separation of the oats paddock from the wheat paddocks. Figures 3(a) and 3(b) show plots of band 5 and of band

7 against time for a selection of paddocks.

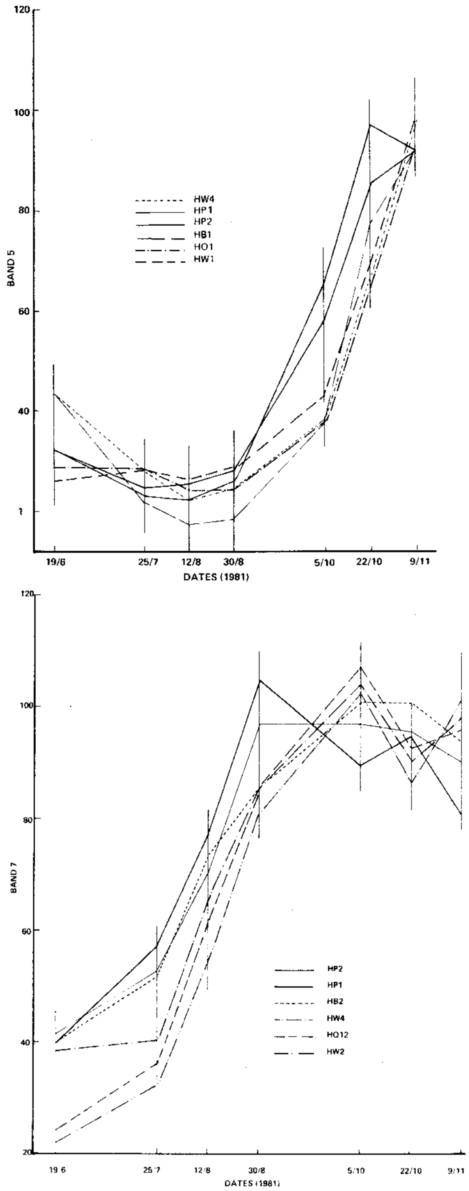

Figure 3. Plots of (a) Band 5 and (b) Band 7 Against Time for a Selection of the Paddocks in Figure 2.

It is of particular interest to CBH to be able to estimate accurately the area sown to crop as early as possible in the season. To this end, analyses have been carried out for overpasses on 25 July and 12 August, and for overpasses on 25 July, 12 August and 29 August; the overpass on June 19 has been excluded since not all areas were sown by this date. Both analyses show reduced separation of crops from pasture, with first two canonical roots of 10.15 and 1.25 and 11.14 and 0.41 respectively. Figure 4 shows the ordination of paddocks based on the overpasses for 25 July and 12 August. Again separation of pasture from crops is evident, though separation among the crop paddocks is markedly reduced.

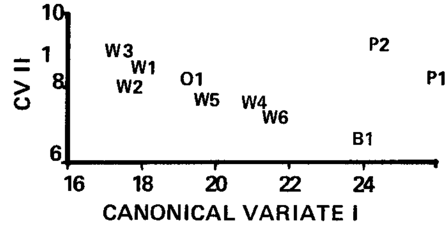

Figure 4. Paddock Means for First Two Canonical Variates for Data from 25 July and 12 August. The canonical vectors are scaled to unit standard deviation within groups.

Even with this reduced separation, the allocation of individual pixels gave very encouraging results. Table 2 summarizes the results of the allocations with the group membership probabilities being pooled for all wheat paddocks and for all pasture paddocks.

Table 2. Results of Allocation of Each Pixel for Each of the Training Paddocks - Leave-One-Out Calculations are Used.

training paddock	no pixels	number of pixels allocated to: W1+...+W6	B1	O1	P1+P2
W1	119	118#	0	2	0
W2	123	121	0	1	0
W3	47	47	0	0	0
W4	53	47	0	6	0
W5	58	48	0	10	0
W6	169	158	5	3	3
B1	47	0	46	0	1
O1	23	23	0	0	0
P1	269	7	5	0	257
P2	199	4	18	0	177

for W1, 118 of the 119 pixels are allocated to one of the wheat groups W1-W6 if a forced allocation is adopted, while 2 pixels are allocated to the oats group.

The results in Table 2 are for forced allocation, with the pixel being allocated to the group or set of groups with the highest group membership probability. An ideal scheme would also incorporate regions of doubt. Most of the pixels are correctly allocated, with very little overlap between crops and pasture. The results for the oats paddock are slightly misleading in that approximately half the pixels are more likely to belong to oats than to any particular wheat paddock, though collectively one or other of the wheat paddocks is more likely.

Tables 3 and 4 summarize pixels that are wrongly allocated or are found to be atypical. In each case, distinct clusterings of pixels are evident. One of the priorities in the field visits this year will be to examine the pasture sites to see if the reasons for these anomalies can be identified.

Table 3. Representation of (a) Pixels Wrongly Allocated Using a Forced Allocation Approach and (b) Pixels Found to be Atypical (at the 0.05 level) for Group P1.

a) misallocated pixels

row \ column	3	4	5	6
1		x	x	
2	x	x	x	x
3			x	
4			x	

b) pixels allocated as pasture but atypical

row \ column	21	22	23	24	25	26	27
5				x	x		
6	x			x			
7						x	
8		x					
9							x
10							x
11				x			x

Table 4. Representation of (a) Pixels Wrongly Allocated Using a Forced Allocation Approach and (b) Pixels Found to be Atypical (at the 0.05 level) for Group P2.

a) misallocated pixels

row \ column	8	9	10	11	12	13
2	x	x	x			
3		x	x			
4				x		x
5					x	
6		x				

b) pixels allocated as pasture but atypical

row \ column	1	2	3	4	5
6	x	x	x	x	
7	x	x		x	
8	x	x	x	x	x
9			x	x	
10			x		

B. RESULTS FOR ADDITIONAL PADDOCKS

A canonical variate analysis of the amended data set based on all four bands for all seven overpasses shows clear and marked separation between all paddocks tentatively labelled as pasture and those labelled as crop (or known to be crop) (Figure 5). The first two canonical roots are 11.91 and 3.28.

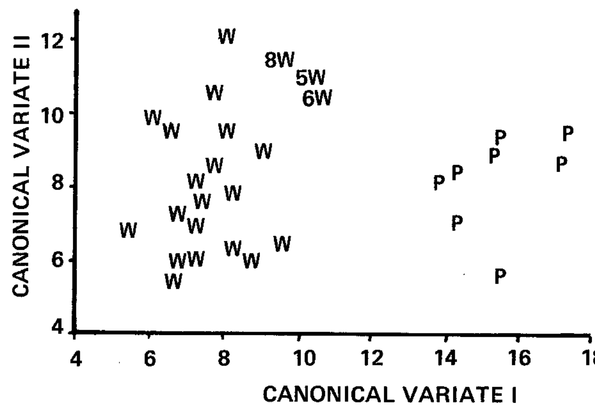

Figure 5. Means for First Two Canonical Variates for All Bands for All Dates. The canonical vectors are scaled to unit standard deviation within groups.

An analysis based on band 5 and band 7 gives corresponding canonical roots of 10.93 and 3.02, so that most of the information for discrimination is contained in these two bands; the last date contributes little to the discrimination. An analysis based on the linear combination (band7-band5) for the first six overpasses gives corresponding canonical roots of 8.56 and 2.46.

A more detailed analysis of the linear combinations (band7-band5) and (band5+band7) shows that the differences for 19 June, 25 July, 12 August, 29 August and 4 October and the sums for 12 August, 29 August, 4 October and

22 October explain much of the separation (with canonical roots of 10.83 and 2.60).

There is obvious separation between pasture paddocks and crop paddocks for (band7-band5) for the first four overpasses (Figure 6).

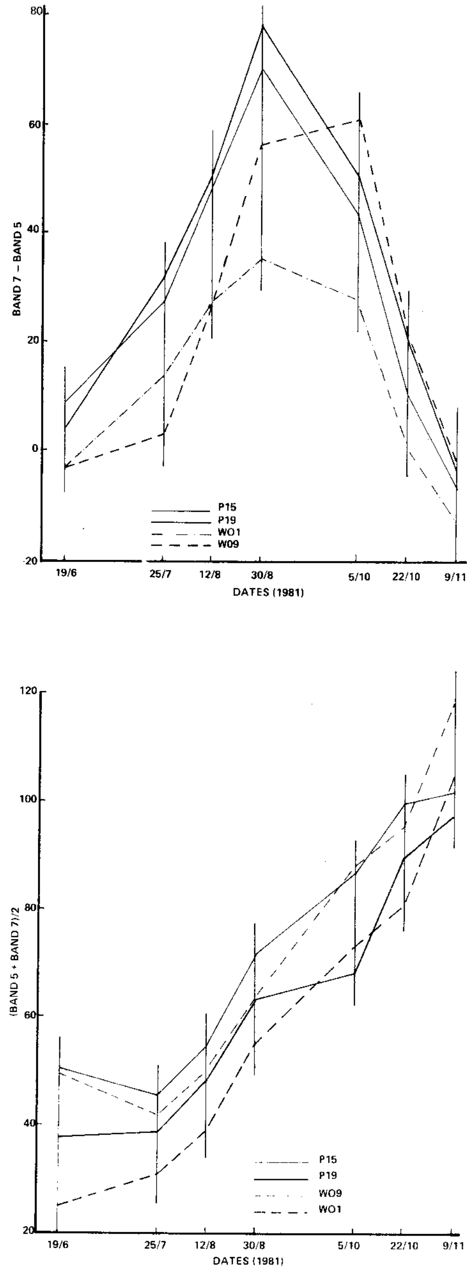

Figure 6. Plots of (a) Band7-Band5 and (b) (Band5+Band7)/2 against Time for a Selection of the Paddocks in Figure 5.

Detailed discussion of separation between the crop paddocks is not possible at this stage. However it is worth noting the behaviour of the three paddocks indicated as 5W, 6W and 8W in Figure 5 (see also one of these in Figure 6). These paddocks were identified at the beginning of this second series of analyses as crop on a light sandplain soil. Their spectral signatures behave quite distinctly on the plot of (band5+band7) against time (Figure 6(b)). This is reflected in the canonical variate plot in Figure 5; these paddocks also show further separation along the third canonical variate.

A robust canonical variate analysis shows a 16% increase in the magnitudes of the canonical roots, with the ordinations being similar for both the usual and robust analyses. Table 5 shows the weights associated with atypical pixels for two paddocks.

Table 5. Non-unit Weights Associated with a Robust Canonical Variate Analysis for Two of the Paddocks.

paddock dimension (in pixels)		column		
		10	11	12
12 rows x 20 columns	row 7	0.02	0.00	0.55
		column		
		1	2	3
10 rows x 21 columns	row 8	0.05	0.09	
	9	0.21	0.12	
	10	0.62	0.15	0.30

The first example shows the effect of a dam in the middle of the paddock, while the second example reflects the influence of a waterlogged area in the corner of the paddock.

The plots for the individual bands in Figure 3 and the sum and difference in Figure 6 show reasonably smooth behaviour. An attempt has been made to see if this can be modelled by an appropriate growth curve. Two approaches have been tried to date. In the first, the position and piecewise slopes of the trends for band 5 and band 7 have been used in place of the original data. Specifically, for band 5, the mean of overpasses on 25 July, 12 August and 30 August is used to define position, with the differences between the mean and the value on 19 June, the mean and the value on 5 October, the values on 5 October and

22 October, and the values on 22 October and 9 November defining the slopes. For band 7, the mean of overpasses on 5 October, 22 October and 9 November is used to define position, with the differences between the mean and the value on 30 August, the values on 30 August and 12 August, the values on 12 August and 25 July, and the values on 25 July and 19 June defining the slopes. The first three canonical roots for the analysis based on bands 5 and 7 are 10.93, 3.02 and 2.22 while those for the analysis based on slope and position are 9.77, 3.00 and 2.13. Further work is underway to refine this description further.

In a second approach, cubic orthogonal polynomials were used to describe the response for (band 7-band 5). The first two canonical roots based on (band7-band5) are 8.56 and 2.46, with substantial separation of crops and pasture. However the analysis based on the four polynomial coefficients (mean, linear, quadratic, cubic) shows considerably less separation, with first two canonical roots of 5.83 and 1.69. The clear separation between crop and pasture is no longer evident.

VII. DISCUSSION

Significant separation between pasture and paddocks sown to crop has been established using a temporal analysis of the data. The allocation of individual pixels to either crop or to pasture is very successful, even for two overpasses, and without using the spatial information in the data.

There are considerable differences between wheat paddocks sown at different times and/or on different soils, and these differences appear to be greater than those between wheat and oats. However results for barley are more encouraging. Detailed paddock histories for farms falling within areas H, F and I are being gathered, to allow this aspect to be examined in more detail.

Within-paddock heterogeneity may cause problems, as shown by the allocation results in Tables 3 and 4 and the robustness results in Table 5. Both the robust calculations and the typicality index give insight into atypical pixels, with both approaches using the individual squared Mahalanobis distance as the basic measure on which to judge typicality or otherwise. How this information can best be incorporated into more routine identification over large areas has yet to be determined.

VIII.. REFERENCES

1. Aitchison, J. and Dunsmore, I.R. (1975). Statistical Prediction Analysis. Cambridge University Press.

2. Aitchison, J., Habbema, J.D.F. and Kay, J.W. (1977). A critical comparison of two methods of statistical discrimination. Appl. Statist., 26, 15-25.

3. Anderson, T.W. (1951). An Introduction to Multivariate Statistical Analysis. Wiley.

4. Campbell, N.A. (1980). Robust procedures in multivariate analysis. I Robust covariance estimation. Appl. Statist., 29, 231-237.

5. Campbell, N.A. (1982). Robust procedures in multivariate analysis. II Robust canonical variate analysis. Appl. Statist., 31.

6. Dawbin, K.W. and Beach, D.W. (1981). Crop Monitoring in Australia Using Digital Analysis of LANDSAT Data. Seventh International Symposium on Machine Processing of Remotely Sensed Data, LARS/Purdue University.

7. McKay, R.J. and Campbell, N.A. (1982). Variable selection techniques in discriminant analysis. I Description. II Allocation. Brit. J. Statist. Math. Psych.

8. Murray, G.D. (1977). A note on the estimation of probability density functions. Biometrika, 64, 150-152.

9. Northcote, K.H. (1971). A Factual Key for Recognition of Australian soils. Rellim Technical Publications.

10. Northcote, K.H., Bettenay, E., Churchward, H.M. and McArthur, W.M. (1967). Atlas of Australian Soils Sheet 5. Melbourne University Press.

IX. ACKNOWLEDGEMENTS

This project would not have been possible without the financial and technical support of CBH. The authors wish to thank T. Dixon and G. Osborne for their assistance, and especially the CBH superintendants and supervisors for their efforts in collecting high quality field data. Thanks are also due to the individual farmers involved in the study.

X. AUTHOR BIOGRAPHICAL DATA

N.A. Campbell is a statistician with CSIRO in Western Australia. He has been involved with the application of multivariate techniques and the development of methodology since 1970.

F.R. Honey is leader of the remote sensing group in CSIRO in Western Australia. He has been involved in the development and application of remote sensing techniques since 1974.

P.T. Hick is a member of the remote sensing group in CSIRO in Western Australia. He has been involved in the application of technology developed by the group since 1975.

M.D.W. Carlton is a temporary member of the remote sensing group in CSIRO in Western Australia. He has been responsible for the development of the programs for the extraction of the Landsat data.