Multispectral Data Analysis:
A Moderate Dimension Example©

David Landgrebe
School of Electrical Engineering
Purdue University
West Lafayette IN 47907-1285
landgreb@ecn.purdue.edu

In this monograph we illustrate the analysis of a multispectral data set of moderate dimensionality, providing results for a range of different processor and parameter selections for the MultiSpec© system. It is assumed that the reader is familiar with MultiSpec©, its various processors, and its documentation entitled “An Introduction to MultiSpec©,” as these will be referred to in describing the analysis. Further, this analysis will rely upon the concepts and principles outlined in the monograph by the author entitled “Multispectral Data Analysis: A Signal Theory Perspective.”

The Data Set.

Flightline C1 (FLC1), a historically significant data set, is located in the southern part of Tippecanoe County, Indiana. It follows a county road from the Grandville Bridge over the Wabash River just south of South River Road (West Lafayette) to near State Highway 25. Though collected with an airborne scanner in June 1966, this data remains contemporary. Key attributes that make it valuable, especially for illustrative purposes, are that it has more than a few spectral bands (12 bands), contains a significant number of vegetative species or ground cover classes (at least 9), includes many regions (e.g., fields) containing a large numbers of contiguous pixels from a given class (thus facilitating quantitative results evaluation), and has "ground truth" available .

The spectral bands in the data set are,

Band #	Wavelength, μm
1	0.40 - 0.44
2	0.44 - 0.46
3	0.46 - 0.48
4	0.48 - 0.50
5	0.50 - 0.52
6	0.52 - 0.55

Band #	Wavelength, μm
7	0.55 - 0.58
8	0.58 - 0.62
9	0.62 - 0.66
10	0.66 - 0.72
11	0.72 - 0.80
12	0.80 - 1.00

The data set consists of 949 scan lines with 220 pixels per scan line, or 208,780 pixels. The scanner used had an instantaneous field of view (IFOV) of 3 milliradians and was flown at an altitude of 2600 ft above terrain. The sensor scans approximately ± 40° about nadir with somewhat less than that being digitized. Each pixel was digitized to 8-bit precision.

An accompanying sheet attached to the end of this document provides ground cover information for each field, according to the following symbols:

Symbol	Species
A	Alfalfa
B.S.	Bare Soil
C	Corn
FS	Farmstead
H	Hay
O	Oats

Symbol	Species
P	Pasture
R	Rye
RC	Red Clover
S	Soybeans
Tim	Timothy(Hay)
W	Wheat

The numbers sometimes marked on fields, e.g. 40" and 90, indicate the canopy height and the per cent ground cover, respectively.

The left segment of the sheet is the north end of the flightline, the right segment the south. There is a small section of the flightline between the printed segments which is not shown. The river in the upper left corner is the only water in the flightline; only the first few scan lines of the data set at the extreme upper left corner contain a portion of the river.

Procedure

1. Using the training set for FLC1 listed in Table 1, run the Feature Selection processor for $1 \leq N \leq 12$. Order the feature sets according to the largest minimum Bhattacharyya distance. Choose the first-listed feature set in each case. The feature sets thus found are given in Table 2.

	Field name	Class	First Line	Last Line	First Col.	Last Col.	No. of Samples
1	Alfalfa1	1	731	737	129	177	343
2	Alfalfa2	1	749	755	131	171	287
3	Alfalfa3	1	809	817	155	183	261
4	Soil1	2	97	119	49	85	851
5	Corn1	3	167	177	33	77	495
6	Corn2	3	267	283	45	61	289
7	Corn3	3	319	341	21	31	253
8	Corn4	3	603	625	13	33	483
9	Oats1	4	421	455	63	83	735
10	Oats2	4	591	599	135	181	423
11	RedCl1	5	439	447	139	183	405
12	RedCl2	5	539	565	175	195	567
13	RedCl3	5	599	619	69	95	567
14	Rye1	6	527	569	127	155	1247
15	Soy1	7	65	81	69	89	357
16	Soy2	7	237	253	141	167	459
17	Soy3	7	307	327	59	81	483
18	Soy4	7	773	777	135	179	225
19	Field69	8	3	4			43
			1	1			
			2	13			
			9	1			
20	Wheat1	9	295	303	134	175	378
21	Wheat2	9	471	495	172	201	750
22	Wheat3	9	607	665	203	211	531
23	Wheat-2-1	10	655	695	17	41	1025
Total Number of training samples							11,457

Table 1. Training Fields of the Standard Training Set.

1	10
2	1,9
3	1,6,9
4	1,6,9,10
5	1,6,9,10,12
6	1,2,6,9,10,12
7	1,2,6,8,9,10,12
8	1,2,6,8,9,10,11,12
9	1,2,3,6,8,9,10,11,12
10	1,2,3,4,6,8,9,10,11,12
11	1,2,3,4,5,6,8,9,10,11,12
12	All 12

Table 2. Feature sets found using the Feature Selection Processor.

2. Classify the flightline using the Maximum Likelihood algorithm with each of these feature sets and determine the training and test sample accuracy, using the standard training and test sets listed in Tables 1 and 3.

Field No.	First Line	Last Line	First Col.	Last Col.	Class	No. of Samples
1	57	89	47	103	Soybeans	1881
2	63	74	115	169	Soybeans	660
3	93	101	113	183	Soybeans	639
4	123	133	43	101	Soybeans	649
5	133	149	43	83	Soybeans	697
6	217	273	109	201	Soybeans	5301
7	705	797	69	111	Soybeans	3999
8	291	341	43	92	Soybeans	2550
9	489	519	115	161	Soybeans	1457
10	643	663	125	197	Soybeans	1533
11	647	659	51	87	Soybeans	481
12	647	675	93	111	Soybeans	551
13	705	797	33	61	Soybeans	2697
14	759	785	121	197	Soybeans	2079
15	157	187	17	101	Corn	2635
16	189	215	17	79	Corn	1701
17	221	255	39	55	Corn	595
18	261	287	39	65	Corn	729
19	307	349	14	35	Corn	946
20	401	421	111	194	Corn	1764
21	589	643	3	43	Corn	2255
22	327	335	109	197	Oats	801
23	365	377	131	185	Oats	715
24	413	467	45	91	Oats	2585
25	583	605	121	191	Oats	1633
26	285	317	109	199	Wheat	3003
27	347	353	107	205	Wheat	693
28	385	393	109	203	Wheat	855
29	459	509	167	211	Wheat	2295
30	581	689	203	211	Wheat	981
31	649	699	3	43	Wheat	2091
32	129	133	113	199	Red Clover	435
33	357	399	61	95	Red Clover	1505
34	433	453	113	197	Red Clover	1785
35	521	561	173	215	Red Clover	1763
36	559	581	49	109	Red Clover	1403
37	589	633	49	109	Red Clover	2745
38	613	619	121	183	Red Clover	441
39	629	637	123	191	Red Clover	621

40	675	695	127	195	Red Clover	1449
41	729	737	121	195	Alfalfa	675
42	745	757	121	195	Alfalfa	975
43	793	815	121	195	Alfalfa	1725
44	525	577	119	163	Rye	2385
45	137	149	87	101	Bare Soil	195
46	95	117	45	89	Bare Soil	1035
					Total	70588

Table 3. Test Set for FLC1.

The results of this Maximum Likelihood pixel classification in terms of the training and test set accuracies are shown in Figure 1. Note that the training set contains 11,457 samples with no class smaller than 851 samples, with the exception of the class Water, which is very small and very spectrally distinct. Thus, no Hughes effect is evident. Note also that the test set contains 70,588 samples, almost seven times the size of the training set, and nearly one third of the total of 208,780 pixels of the entire flightline.

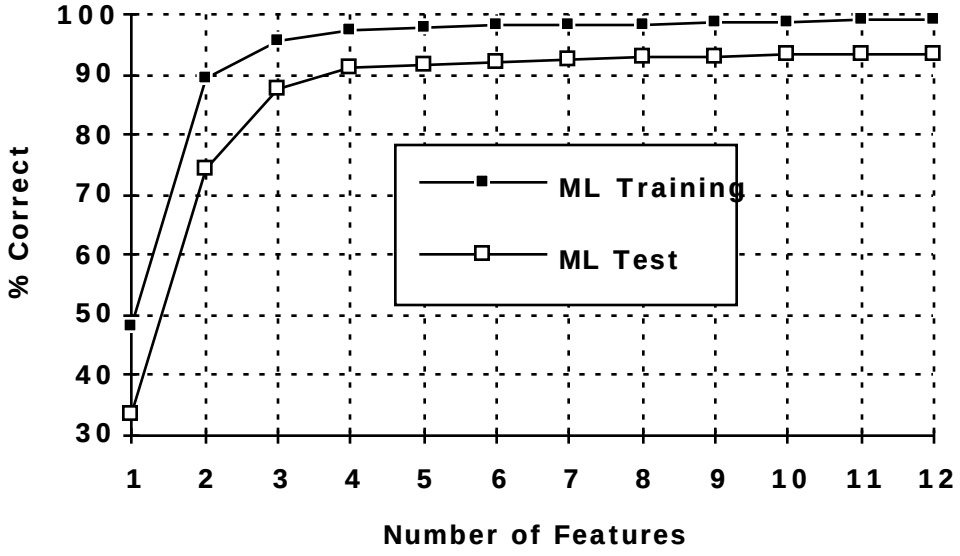


Figure 1. FLC1 Training and Test Sample Classification Accuracy

3. Run the DBFE Feature Extraction algorithm using the same training set. Create the transformed data set from the resulting transformation matrix using the Reformat processor. Classify the flightline using the Maximum Likelihood pixel classifier with this transformed data set using feature set sizes of the first 1 through the first 12. Determine the training set and test set accuracies for each. The results for test sample accuracy compared to that for the original, untransformed bands are shown in Figure 2. The results for the DBFE case are seen to be slightly better, but the small improvement may not be statistically significant. The DBFE algorithm has the advantage that its output does provide information about how many features to use in making the classification. It also functions very satisfactorily for data sets of much larger dimensionality.

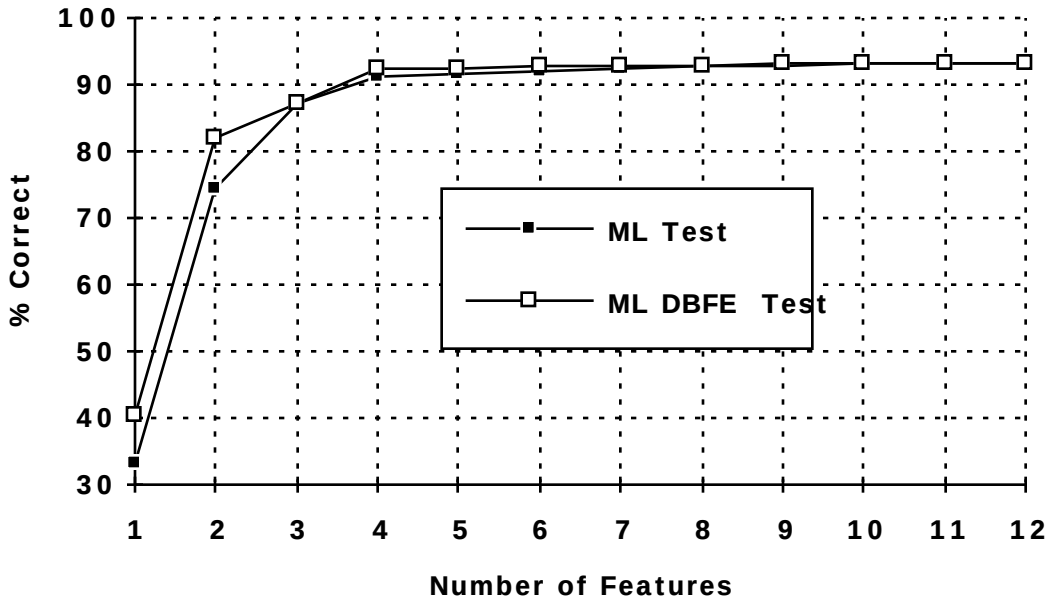


Figure 2. Accuracy for Spectral Bands vs. DBFE Features

4. To improve the performance further, classify the flightline with DBFE features using the ECHO spectral/spatial classifier. The results are compared with the Maximum Likelihood pixel classification in Figure 3.

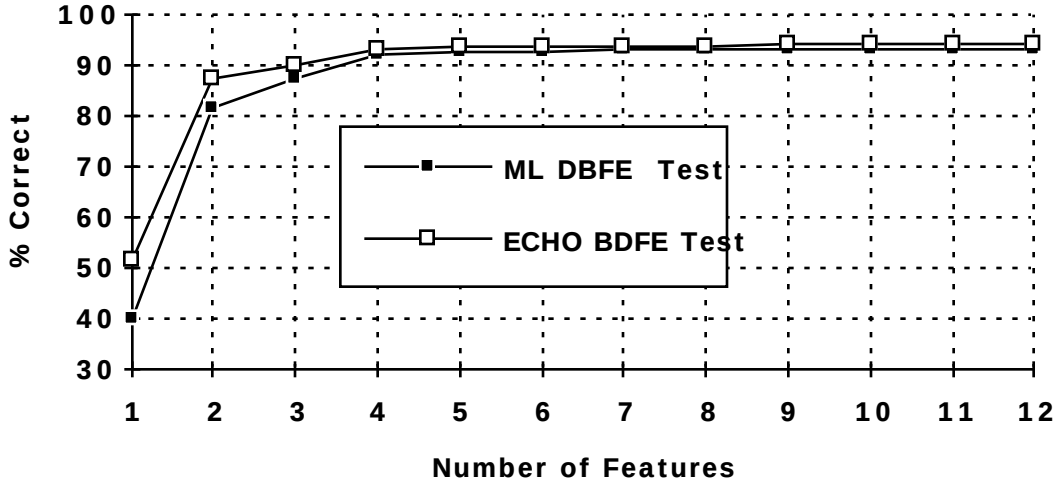


Figure 3. ECHO vs. Maximum Likelihood results using the DBFE features.

It is seen in these tests that use of DBFE over the original spectral bands and ECHO over Maximum Likelihood pixel classification improves performance only marginally but consistently. Table 4 summarizes all results to this point in tabular form. In a sense, this data set does not provide a severe enough test of the marginal improvement which these techniques may generally be expected to provide, because the classes are relatively separable. Only three or four features are required to achieve 90% accuracy or above.

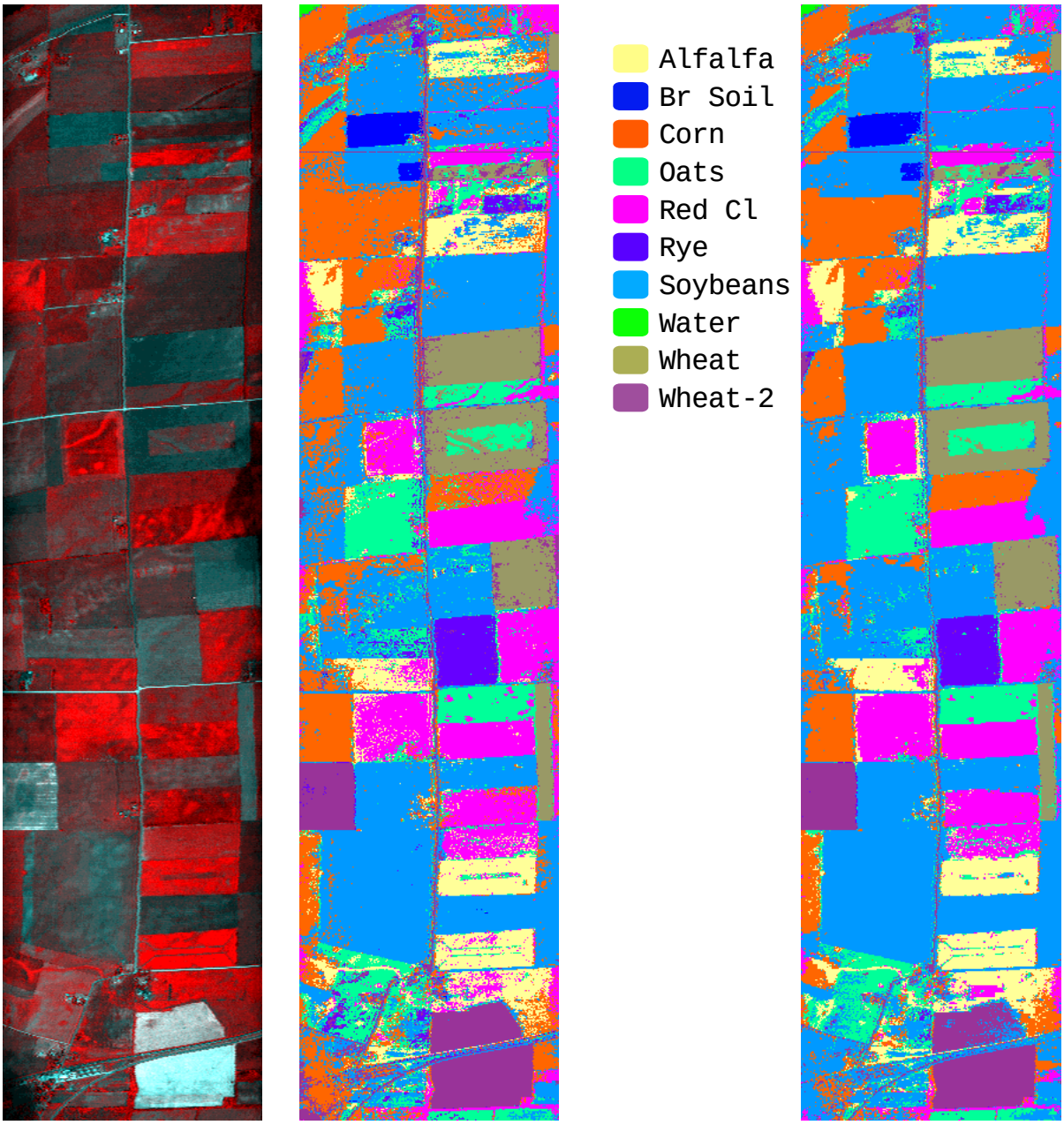
A key characteristic of any such analysis is the generalization capability of the classifier. That is, how well does the classifier perform on samples other than its training samples, and how well does this generalization capability hold up over the entire data set. The use of test fields was designed to measure this characteristic to the extent that it can be done so quantitatively.

However, this generalization characteristic is very dependent upon the analyst's selection of the training set. To assist in this process, a processor called Enhance Statistics is contained in MultiSpec. It iteratively adjusts the training statistics using a combination of the original training statistics and a uniform sampling of the entire data set to achieve class models with maximum class likelihood over the entire data set. As might be suspected, this may sometimes result in a modest decline in the measured training set accuracy or even the broader test set, but should produce a better overall analysis.

5. Run the Enhance Statistics processor on the DBFE transformed data set and classify the data set using the resulting enhanced statistics. Either the Maximum Likelihood pixel classifier or the ECHO algorithm may be used. The right hand four columns of Table 4 provide some results of this process. It is seen that in this case, the quantitatively measurable accuracies remain at their previous levels or improve slightly. Figure 4 compares the data in image form with the enhance statistics Maximum Likelihood and the enhanced statistics ECHO result.

	Max. Lik.		Max. Lik. DBFE		ECHO DBFE		Enhanced Max. Lik		Enhanced ECHO	
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
1	48	33	68	40	80	52				
2	89	75	92	82	99	87				
3	96	88	95	87	99	90				
4	97	91	98	93	99	94	98	93	99	95
5	98	92	99	93	100	94				
6	98	92	99	93	100	94				
7	98	93	99	93	100	94				
8	99	93	99	93	100	94				
9	99	93	99	93	100	94				
10	99	93	99	93	100	94				
11	99	93	99	93	100	94				
12	99	93	99	93	100	94	98	93	100	96

Table 4. Summary of all results.



A. Simulated CIR Image (in color)

B. 12 Band Max. Likeli. Classification (in color)

C. 12 Feature ECHO Classification using Enhanced Statistics (in color)

Figure 4. The data set in image form and results from analysis illustrating the effects of enhanced statistics.

One of the ways to quantitatively estimate the effect of the Enhance Statistics processor is to compare before and after values of the Average Likelihood Probability, which are provided by MultiSpec whenever a Probability Results File is created during a classification. Table 5 lists results showing a substantial increase in the Average Likelihood Probability from its initial value for the Maximum Likelihood classification to the final Enhanced ECHO analysis as a result of the Statistics Enhancement Process. This table also shows another frequently observed advantage of the ECHO processor. Not only does it tend to improve accuracy over Maximum Likelihood pixel classification, but it often provides a significant reduction in processing time.

No. of Features	Ave. Prob. % ML	Enh.Ec	Proc. Time ML	ECHO
4	19	36	202	140
12	21	35	537	358

Table 5. Average Likelihood and processing time for several parameter settings.

Concluding Remarks

This document is intended to point to some of the possible processing steps in analysis of a multispectral data set and illustrate their likely impact in a typical situation. Further, enough information is provided so that the reader, after acquiring a copy of MultiSpec, its documentation, and the FLC1 data set could reproduce the results obtained and try other options.

It is clear from such exercises that, no matter what algorithms are used to analyze multispectral data, the aspect of greatest importance is the accurate and thorough modeling of the classes of interest, relative to the other spectral responses that exist in the data set, and doing so in such a manner as to maximize the separability between them in feature space. This fact places great emphasis upon the analyst and his/her skill and knowledge about the scene. The current goal of the on-going research effort in this field thus is to find ever improving tools to assist in this process and increase its objectivity.