HANUSCHAK

78-13

THE 1976 ILLINOIS PROJECT FOLLOW-ON STUDY

by Paul W. Cook

1978

Introduction

Many questions concerning the use of LANDSAT data in making corn and soybean crop hectarage estimates were answered during the 1975 Illinois Project [1]. The success of this project, however, caused the question to be raised as to whether or not such results could be expected each year if an operational system were implemented. As a means of examining this problem, data for the 1976 growing season was collected to provide the necessary ground data to work with the LANDSAT data obtained during the summer months.

By the end of the growing season of 1976, the Illinois Weekly Weather and Crop Bulletin indicated that 92% of the corn had reached the dent stage by September 12, 1976, and that 77% of the soybeans had turned yellow while 45% were shedding leaves. The summer of 1975 was very much similar since by September 2, 1975, 90% of corn was in dent and 55% of the soybeans had turned yellow.

The previous four-year average for 1970-1974 had 62% of corn in dent stage by September 2, while only 14% of the soybeans were turning yellow. This early maturity of the crops indicates that 1976 was more similar to the 1975 growing season than to that of the previous four-year average.

Since the crop maturity was more like that of 1975, our experience during that analysis showed that August LANDSAT data would be the most desirable. However, cloudy weather occurred during June, July, and August for most of the state. September brought clear skies and so the entire state was cloud free for the nine LANDSAT scenes necessary for full coverage on September 11th and 12th.

The availability of cloud-free data left little choice as to what time of year would be best to do the data analysis. Results for September 1975 had been much worse than for July or August. Use of September 1976 data, therefore, could be of at best questionable value, but worth at least an examination of one strip to determine if September imagery would be of more value during 1976.

Data Preparation

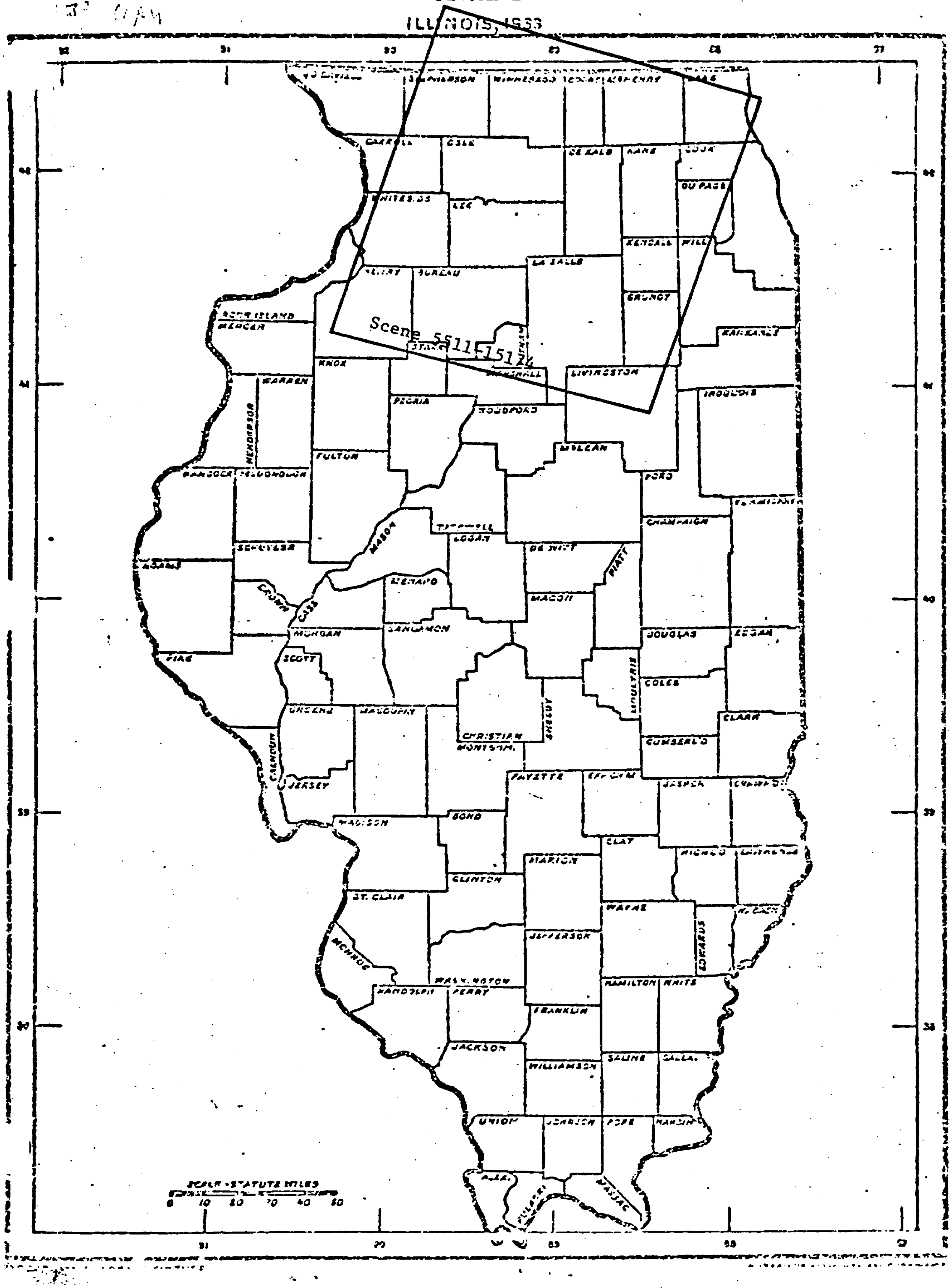
Three scenes from one strip were chosen from the available September data. They were scenes 5511-15174, 5511-15181, and 5511-15183 from September 11, 1976. All scenes were LANDSAT I and covered a swath of Illinois from the north-eastern ge near Chicago to the south-western border along the Mississippi, Missouri, and Ohio River confluence (see Figure 1). Only one scene was analyzed, however, and so only it is shown.

Unlike previous work in Illinois, the LANDSAT scenes were not subjected to a skew-correction algorithm. Instead, each tape was reformatted to BBN format in such a way that the original skew of the sensor was maintained and each pixel was shifted by a constant number of rows and columns (166 rows and 373 columns). This shift was caused by the program normally used to deskew and reformat LANDSAT tapes in one operation.

One step was eliminated in this process. No longer was it necessary to choose points on the photo and USGS maps, digitize these point pairs, and run a first order equation for determination of the deskewing parameters. Instead, the parameters were chosen as zero and the program run after the LANDSAT tapes were received.

FIGURE 1

ILLINOIS, 1933



Another change in the registration procedure involved the selection of fewer registration points in an effort to determine what number of control points would be optimal for accurate registration of a LANDSAT scene. Thirty-two points (rather than the usual sixty or more) well scattered throughout the LANDSAT were chosen at this point. Since the tape was already in BBN format and at BBN, grey-scales could be printed immediately after point selection on the photo had been accomplished.

Maps for each point were also selected and the grey-scale overlayed. An appropriate pixel center was chosen within each grey-scale for a corresponding point on the map. Each corresponding map point was digitized and the pixel's coordinates entered into a file for evaluation by both a full third-order polynomial and DAM-COEFF [2]. After correction of inaccurately located points, the final registration accuracy achieved was as follows (note: all thirty-two points were used without letions):

	Pixel Errors		RMS		
	Max Line	Max Column	Line	Column	Meters
3rd Order Polynomial	-1.40	0.44	0.47	0.62	51.4
DAM-COEFF	1.10	4.36	0.68	1.85	118.5

After creation of a third-order calibration file of the appropriate coefficients needed in predicting (line, column) locations within the LANDSAT digital data, masks for each of the eighty-one segments located on the scene were made from the previously digitized segment network files. At the same time, a coordinates file for each segment was created and used to determine the area for printing out grey-scale maps about each segment.

As a means of determining the proper location for each segment, a plot was made showing the location of the segment in terms of LANDSAT row and column as well as belling each tract and field within the segment. Lightness and darkness ttern within the computer grey-scale corresponding to fields of different crops (generally either corn or soybeans) were used in conjunction with the plot to determine the correct location for each segment. By shifting the plot on the grey-scale, one could determine the correct placement and thereby the amount of shift necessary in both row and column (see figure 2).

Each segment shift was digitized and a file of the shifts generated (see figure 3). Local calibration files for each moved segment was also made. This allowed creation of segment mask files to be used in extracting LANDSAT data for analysis.

Indications of some spurious data being present caused a further step in segment location to be performed. Data for corn, soybeans, and all other crops were clustered individually and the resulting statistics file used to classify all the pixels within each segment window.

Each window was printed out and the categories named C (corn), S (soybeans), ᴉd O (other) for each pixel. Placing the segment plots on these print-outs allowed an additional check of correct placement by using concentrations of cluster-types corresponding to the crops of interest - (corn or soybeans). This procedure resulted in further movement of some segments (see Figure 4).

All movements were made relative to the locations predicted by the global calibration file. No movement was greater than 2.09 pixels in row nor 5.88 pixels in column. Overall errors were 0.95 pixel rms in row and 2.63 pixels rms in column. Since some of this error of segment location may also be ascribed to errors in calibration of the segment photos to map, segment location seemed to be reasonably accurate using the calibration file generated by only 32 points. Of eighty-one segments on the scene, only forty-six were found to need movement using this technique. Naturally, if the additional unshifted segments were also considered, then the rms shift values would be even further reduced.

Data Analysis

When it had been determined that the segments were correctly located, further analysis of the data began. Because eight major crop categories were present in the data, the first attempt was made to determine if this natural grouping was also spectrally separable.

Using the minus cover option in the data packing program, only LANDSAT pixels within the interior of each field of a given crop were extracted from files and combined to provide a file of each crop type. Means and variance-covariance matrices were calculated for each data file and the resulting eight files combined into one statistics file for further use.

Meanwhile, another file with all crop pixels, including field boundary pixels, was also packed (called not-background or NB). This file of over 35,000 pixels would be used to test each classifier as to per cent correct and for small scale estimation checks of the correlation between classified pixels to reported acreage for each crop type. The ratio of the direct expansion estimate's variance to that obtained from single regression estimation was called the RE_2 value and was used as a further comparison of how well a given statistics file had performed.

The not-background file was used to test this first statistics file consisting of eight cover types with one cluster per cover type. Results of this analysis were not promising and seemed to confirm the initial observation of the crops not being separable as had been indicated by the Swain-Fu distances between the clusters.

The usual method of improving the separability of the statistics for given crop types is to cluster each given crop category into a multitude of small cluster groups. Variances for each cluster group are smaller and any non-normal data structure is more closely modeled by a larger number of clusters used to follow the size and shape of the actual data structure.

Since corn and soybeans were of primary interest, all other crop categories within the segments were included into one packed file. The number of pixels found for these files were as follows:

Crop	Approximate number of pixels
Corn	12,000
Soybeans	4,000
All other covers	5,000

Note: 1. All other covers consisted of waste-land, permanent pasture, dense woods, alfalfa, oats, oat and wheat stubble, and other hay.

2. Only interior pixels were included in the packed files.

Each file was then clustered so that a total of thirty-six clusters were created. Again, the NB was classified and the accuracies obtained were checked. This showed little improvement and so each packed crops file was clustered again to obtain a total of sixty groups.

Contrary to expectations, the sixty category classification actually gave lower r^2 and RE_2 results than had the 36 category classification - even though per cent correct for both corn and soybeans had improved. Since more clusters did not seem to be the answer, the next step would be to explore ways of reducing clusters and further improving r^2 and RE_2.

The first method used for simplifying the number of clusters is a program known as Group Categories Automatically. It uses the spectral (or optionally, spatial) information contained in the statistics file to determine which clusters would be most efficiently grouped (see example 1). A cut-off of 10% relative transmission loss was selected as the maximum to be acceptable for each crop statistics file.

The next step after combining categories was to recluster the packed crops files using those categories obtained from combining the original clusters according to Group Categories Automatically. The purpose of this process is the development of new crop clusters which more accurately reflect the properties of the data being analyzed. An improvement in correct classification râtes as well as r^2 values would indicate that this was achieved. See examples 2 and 3 for how the seed clustering was done.

Results of the Group Categories Automatically gave the following reduction in the number of categories:

	Original Clusters	Combined Clusters
(1) Corn	26	10
(2) Soybeans	22	10
(3) Other Crops	12	9
Total	60	29

Each newly created file was then used to seed the clustering for each file. In every case of clustering 100.0% convergence and all steps necessary for convergence was used. Additionally, no cluster combining was allowed by the clustering program since such cluster combining had already been done by using the Group Categories Automatically command.

The three statistics files obtained from the above seeded clusterings were further combined and used to classify the NB data. Both equal priors and priors proportional to expanded reported acreage were used in conjunction with the final 29 clusters obtained.

The best results occurred using the 29 categories with equal priors. For neither corn nor soybeans were the per cent corrects the highest. The r^2 and RE_2, however, were both significantly improved over any of the other methods used for classification. The gain in RE_2 was most significant for soybeans, but even corn showed some marginal improvement (especially in r^2 values) (see Table 1a and 1b).

Even with the improvements as noted from using the clustering method as implemented here, the overall results were not up to the quality that would be needed to do a full state-wide study. It was therefore decided not to pursue an entire study since no better data was available and the estimates would not be sufficiently timely to justify further efforts.

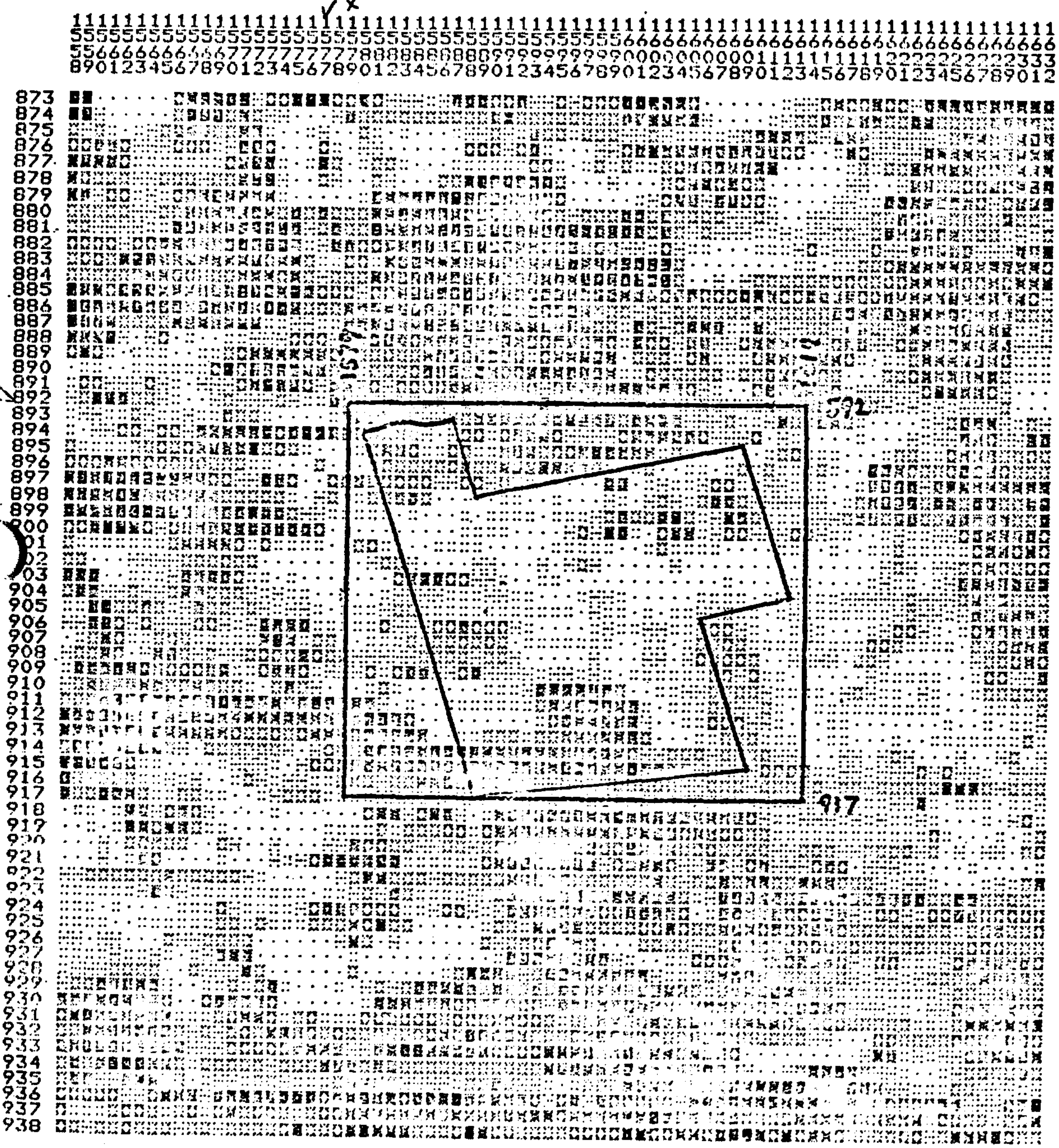
Conclusions

Even though this project did not produce county hectarage estimates for Illinois for the 1976 crop season, two potentially helpful conclusions can be made. The first such conclusion is that the 1975 Illinois Project did establish quite correctly that late season crop detection is not of sufficient quality to allow accurate crop hectarage estimates to be made. Another conclusion to be made is that further testing of methods to do clustering (especially cluster seeding) may be of great value in improving the accuracy of crop hectarage estimation.

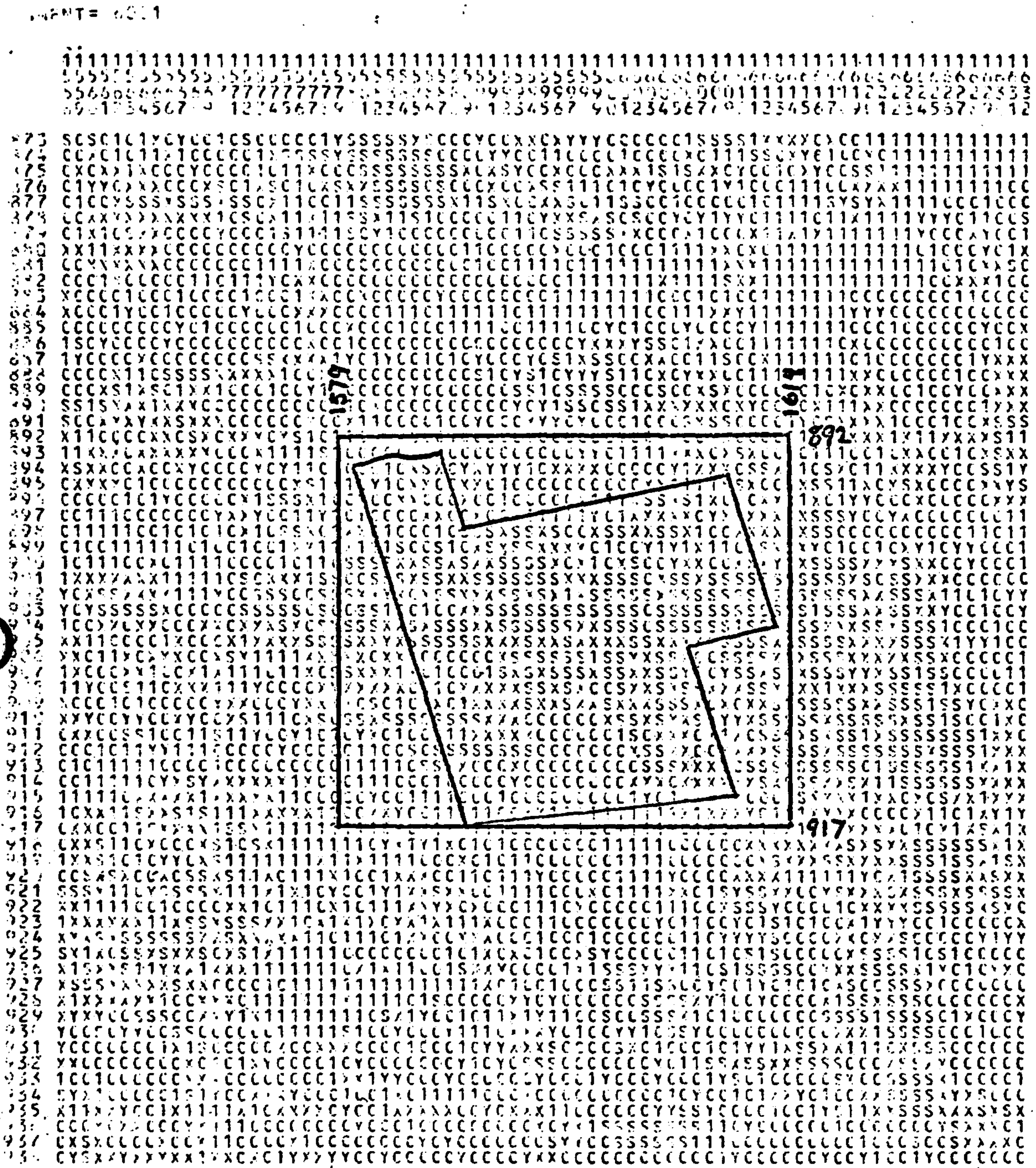
Because insufficient data was available for earlier crop classification (especially during what was the best time period during 1975 - i.e., August), no further confirmation can be made of the 1975 results. Testing these conclusions would of course require a new study.

CHANNEL 2 5511-15174,SEGS INTIAL

SEGMENT= 6001



Segment 6001, Illinois - Grey-scale location. Only the segment boundaries were drawn to simplify the illustration.

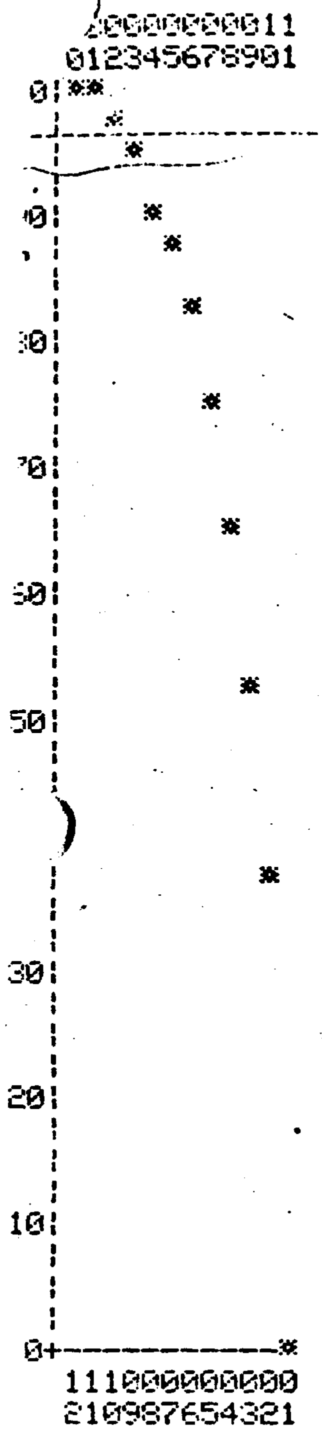


Segment 6001, Illinois

Location of segment 6001 on the classified print-out. Only the segment boundaries were drawn to simplify the illustration.

```
2!GROUP

ENTER PRINT DEVICE:
T=>TERMINAL OR B=>BACKUP DISK T

ENTER "S" TO GROUP WITH SPECTRAL STATISTICS
ENTER "A" TO GROUP WITH ADJACENCIES
#S

ENTER NAME OF INPUT STATISTICS FILE
INPUT STATISTICS FILE=15174.STAT/12-DWOODS


NUMBER OF CATEGORIES= 12 NUMBER OF CHANNELS=4
ENTER "E" FOR EQUAL WEIGHTS OR "D" FOR DISTRIBUTION WEIGHTS E


MINIMUM-RELATIONAL-TRANSMISSION-LOSS
HIERARCHIAL CLUSTER ANALYSIS PROCEDURE:
CHAR 125BC347689A

TENS 000110000001
UNIT 125123476890
 0    125BC347689A
      000110000001
      125123476890
      !!!!!*+!!!!!
 1    125BC347689A
      000110 00001
      125123 76890
      !!!!!! !*+!!
 2    125BC347689A
      000110 00 01
      125123 76 90
      !!!!!! !! *+
 3    125BC347689A
      000110 00 0
      125123 76 9
      !*+!!! !! !
 4    125BC347689A
      00 110 00 0
      12 123 76 9
      !! !!*-+! !
 5    125BC347689A
      00 110  0 0
      12 123  6 9
      !*-+!!  ! !
 6    125BC347689A
      00  10  0 0
      12  23  6 9
      !!  !*--+ !
 7    125BC347689A
      00  10    0
      12  23    9
      *+  !!    !
 8    125BC347689A
      0   10    0
      1   23    9
      !   !*----+
 9    125BC347689A
      0   10
      1   23
      *---+!
10    125BC347689A
      0   0
      1   3
      *---+
```

NUMBER OF STAGE OF PAIRWISE CLUSTER MERGING



NUMBER OF CLUSTERS REMAINING AT EACH STAGE

PATTERN TRANSMISSION LOSS FUNCTION FOR CLUSTER ANALYSIS

(PERCENT PATTERN TRANSMISSION RETAINED)

YOU WISH TO CONTINUE (Y OR N)? N

CPU TIME: 7.40 ELAPSED TIME: 2:45.84

NO EXECUTION ERRORS DETECTED

```
CLUS
JMBER OF CHANNELS=4
HOOSE TYPE OF CLUSTERING:ORD

SE AN INPUT STATISTICS FILE TO INITIALIZE MODE CENTERS (Y OR N)? N

NPUT WINDOW FILE=15174.PACK/-DWOODS

ATEGORIES ASKED=12, DELTA= 0.00, ROW COL SAMPLING= 1 1
HANNELS=4, TOTAL NUMBER OF POINTS=  934


INIMUM NUMBER OF CATEGORIES AFTER MERGING (1 TO 12) 12
HAT PERCENT CONVERGENCE?(0.001-100.0) 100

NTER MAXIMUM NUMBER OF ITERATIONS TO BE PERFORMED
(-1 TO PERFORM AS MANY AS NEEDED TO ATTAIN DESIRED CONVERGENCE)
-1

ERCENT CONVERGENCE=  0.00
ERCENT CONVERGENCE= 47.00
  ENT CONVERGENCE= 74.41
  NT CONVERGENCE= 89.83
ERCENT CONVERGENCE= 91.97
ERCENT CONVERGENCE= 95.29
ERCENT CONVERGENCE= 96.57
ERCENT CONVERGENCE= 97.86
ERCENT CONVERGENCE= 98.72
ERCENT CONVERGENCE= 99.04
ERCENT CONVERGENCE= 98.72
ERCENT CONVERGENCE= 98.29
ERCENT CONVERGENCE= 98.50
 PERCENT,=ONVER#ENCE= 98.39
ERCENT CONVERGENCE= 98.50
ERCENT CONVERGENCE= 99.61
ERCENT CONVERGENCE= 98.50
ERCENT CONVERGENCE= 98.07
ERCENT CONVERGENCE= 98.82
ERCENT CONVERGENCE= 99.04
ERCENT CONVERGENCE= 98.50
ERCENT CONVERGENCE= 99.04
ERCENT CONVERGENCE= 99.57
ERCENT CONVERGENCE= 99.89
ERCENT CONVERGENCE=100.00
25 ITERATION(S).
JMBER OF POINTS IN EACH CLUSTER:
 42   42  122  130   95   78  151   76   93   40
      19
```

```
MAL CLASSIFICATION

MBER OF POINTS IN EACH CLUSTER:
 42    42  122   130    95    78  151    76    93    40
 46    19
 SEE CLASSIFICATION, CREATE A CATEGORIZED WINDOW
LE AND USE THE "PRINT" COMMAND.
NT TO SEE STATISTICS?(Y OR N) Y

                    SEPARABILITY MATRIX

  + 1.00
  + 0.82 1.00
  + 0.76 1.43 1.00
  + 1.19 1.47 0.69 1.00
  + 0.80 0.70 0.82 0.82 1.00
  + 1.23 1.08 1.09 0.64 0.70 1.00
  + 1.49 1.95 0.93 0.63 1.39 0.99 1.00
  + 1.69 1.60 1.51 0.96 1.39 0.72 0.75 1.00
  + 1.87 2.14 1.57 1.19 1.78 1.27 0.67 0.74 1.00
  + 1.86 1.78 1.79 1.39 1.68 1.22 1.09 0.69 0.64 1.00
  + 1.34 0.67 1.80 1.52 1.06 0.88 1.73 0.97 1.65 1.21 1.00
  + 1.59 0.98 2.42 2.30 1.65 1.69 2.57 1.71 2.59 1.94 0.82 1.00
  +  +    +    +    +    +    +    +    +    +    +    +    +
     1    2    3    4    5    6    7    8    9    A    B    C

ROUP               MEANS                            VARIANCES
 1      19.00   12.62   25.60   13.21        1.85    5.17    3.86    3.49
 2      23.29   19.19   30.21   14.71        2.01    3.23    2.71    1.87
 3      18.31   10.95   30.88   17.70        0.68    1.04    1.20    1.57
 4      19.27   12.02   33.58   16.33        0.74    0.77    0.85    1.29
 5      21.08   14.79   31.05   16.21        1.14    1.40    1.18    1.36
 6      21.06   14.82   34.35   18.38        0.94    1.37    1.14    1.15
 7      18.42   10.79   35.30   20.41        0.54    0.90    1.25    0.83
 8      20.96   14.17   37.72   19.67        1.27    2.36    1.16    1.05
 9      18.84   11.00   38.61   21.75        0.77    0.89    1.22    1.43
 A      20.65   13.25   41.87   22.45        1.72    2.65    2.68    1.69
 B      24.28   20.46   35.54   18.09        1.90    2.79    3.85    1.50
 C      28.47   27.68   36.74   18.11        2.60    4.78    5.76    1.99

REATE A CATEGORIZED WINDOW FILE?(Y OR N) N
REATE A STATISTICS FILE?(Y OR N) Y

UTPUT STATISTICS FILE=15174.STAT/12-DWOODS

PU TIME: 1:41.53        ELAPSED TIME: 4:55.37
O EXECUTION ERRORS DETECTED
```

```
CLUS
MBER OF CHANNELS=4
OOSE TYPE OF CLUSTERING:ORD

E AN INPUT STATISTICS FILE TO INITIALIZE MODE CENTERS (Y OR N)? Y

PUT STATISTICS FILE=15174.STAT/8-DW$OODS;1
                                          INPUT STATISTICS FILE TYPE=4
MBER OF CATEGORIES= 8 NUMBER OF CHANNELS=4
PUT WINDOW FILE=15174.PACK/-DWOODS

TEGORIES ASKED= 8, DELTA= 0.00, ROW COL SAMPLING= 1 1
ANNELS=4, TOTAL NUMBER OF POINTS=  934


  UM NUMBER OF CATEGORIES AFTER MERGING (1 TO  8) 8
   PERCENT CONVERGENCE?(0.001-100.0) 100

TER MAXIMUM NUMBER OF ITERATIONS TO BE PERFORMED
-1 TO PERFORM AS MANY AS NEEDED TO ATTAIN DESIRED CONVERGENCE)
-1

RCENT CONVERGENCE=  0.00
RCENT CONVERGENCE= 98.18
RCENT CONVERGENCE= 98.39
RCENT CONVERGENCE= 99.14
RCENT CONVERGENCE= 99.36
RCENT CONVERGENCE= 99.36
RCENT CONVERGENCE= 99.79
RCENT CONVERGENCE=100.00
  8 ITERATION(S).
UMBER OF POINTS IN EACH CLUSTER:
 49  110  206  140  215  138   57   19

INAL CLASSIFICATION

UMBER OF POINTS IN EACH CLUSTER:
 49  110  206  140  215  138   57   19
 SEE CLASSIFICATION, CREATE A CATEGORIZED WINDOW
ILE AND USE THE "PRINT" COMMAND.
 T TO SEE STATISTICS?(Y OR N) Y
```

SEPARABILITY MATRIX

```
  + 1.00------------------------------------------------------------------------
: + 0.75 1.00
3 + 0.74 0.70 1.00
4 + 1.05 0.75 0.67 1.00
5 + 1.20 1.06 0.66 0.73 1.00
5 + 1.50 1.37 1.12 0.76 0.62 1.00
7 + 1.31 0.85 1.43 0.91 1.61 1.27 1.00
8 + 1.65 1.41 2.06 1.68 2.44 2.02 0.81 1.00
  +  +    +    +    +    +    +    +    +
     1    2    3    4    5    6    7    8
```

ROUP	MEANS				VARIANCES			
1	18.92	12.43	26.04	13.61	1.66	4.29	4.54	3.74
2	21.78	16.17	30.64	15.58	1.57	3.10	1.96	2.03
3	18.75	11.50	31.78	17.77	1.03	1.55	1.72	1.38
4	20.96	14.49	35.47	18.75	1.00	1.73	3.11	1.46
5	18.66	11.00	35.27	20.21	0.78	0.99	1.61	1.03
6	19.65	12.09	39.71	21.86	1.98	2.63	3.32	1.73
7	24.30	20.63	34.81	17.65	2.00	2.74	5.62	2.34
8	28.47	27.68	36.74	18.11	2.60	4.78	5.76	1.99

REATE A CATEGORIZED WINDOW FILE?(Y OR N) N
REATE A STATISTICS FILE?(Y OR N) Y

UTPUT STATISTICS FILE=15174.STAT/8-DWOOD2

PU TIME: 29.02 ELAPSED TIME: 3:6.68
0 EXECUTION ERRORS DETECTED

XIT.

TABLE 1a

Corn

Categories	r^2 (Stratum 11, 12)	% Correct	Priors	RE_2
29/MCPC	.1201 .0004	83%	PER	1.01
29/MCPC	.2292 .0013	64%	EQUAL	1.09
8/SCPC	.1741 .0053	54%	EQUAL	1.05
60/MCPC	.1515 .0000	77%	EQUAL	1.03
36/MCPC	.1606 .0013	72%	EQUAL	1.04

TABLE 1b

Soybeans

Categories	r^2 (Stratum 11, 12)	% Correct	Priors	RE_2
(5) 29/MCPC	.3772 .0013	36%	PER	1.41
(4) 29/MCPC	.5258 .1861	43%	EQUAL	1.98
(1) 8/SCPC	.3698 .0203	47%	EQUAL	1.50
(3) 60/MCPC	.3658 .0203	58%	EQUAL	1.49
(2) 36/MCPC	.4330 .0063	48%	EQUAL	1.54

The 8 categories were the following:

1) Corn
2) Soybeans
3) Wasteland
4) Permanent Pasture
5) Dense Woods
6) Alfalfa and Oats
7) Oats and Wheats
8) Other Hay

The 36 clusters were the following:

		Number of Groups
a.	Corn	12
b.	Soybeans	10
c.	Other	14

The 29 clusters wer the following:

		Number of groups
1)	Corn	10
2)	Soybeans	10
3)	Mixture of the other - 6 crops	9

The 29 clusters were obtained from the original 60 clusters as mentioned previously

		Number of groups
1)	Corn	26
2)	Soybeans	22
3)	Others	12