Draft 9/07/2005

Using FSA Administrative Data in the NASS Cropland Data Layer

Michael Craig
Senior Remote Sensing Analyst
USDA/NASS/RDD/GIB/SARS

INTRODUCTION

Both the National Agricultural Statistics Service (NASS) and the Farm Service Agency (FSA) collect field level crop information from US farmers, although with different specifications and goals. Several different approaches to sharing and using collected information are being considered at this time by both agencies [Dorn, 2005]. This specific research started in 2001 with a small area pilot project to look for mutual benefits between the NASS Cropland Data Layer (CDL) Project [Craig, 2001; Hanuschak, 2001] and the FSA Common Land Unit (CLU) Geographic Information System Program. An improved CDL product was felt to be beneficial to both agencies. Some questions to be answered, with respect to CLU use in the NASS CDL Project, are:

> How do FSA and NASS data sets compare over the same areas?
> Can the FSA data alone be used to generate spectral signatures for the CDL?
> Can the FSA data be used to augment existing NASS data for CDL signatures?
> Can extra minor crop information/signatures be generated from the FSA data?
> How do we best convert FSA data into a usable format for the CDL system?

The 2001 joint pilot project consisted of five counties (Gage, Jefferson, Lancaster, Seward, and Saline) in Southeastern Nebraska. Nebraska was selected because it is one of the leaders in the CLU program. The five counties make up a rectangle that is completely contained in a single Landsat scene. In addition, these counties have enough NASS June area segments overall to make a reasonable classification without additional data from the FSA administrative data system. The main summer crops in this region are corn and soybean; although sorghum, winter wheat, and alfalfa are present. The pilot project was completed in September, 2002. Specific results of this project are detailed in Appendix I. Overall, the prospect of using FSA polygons to enhance the Cropland Data Layer Program at NASS was encouraging. However, results were mixed with respect to estimation of the more minor crops.

Based on the promising pilot results, an entire state research effort was done for the 2003 crop season. Nebraska was selected for continued research because it was the most crop intensive state of those which had complete CLU coverage. The 2003 data was delivered with associated crop signup data already populated in the county files. Thanks to the cooperation of the Nebraska state FSA office, additional CLU data was obtained for the 2002 and 2004 crop seasons. Crop season 2004 CLU files were obtained from the Wisconsin and Florida FSA offices under similar cooperative efforts; these two datasets are being used in support of ongoing operational projects.

As mentioned above, other (i.e., not specifically related to the NASS CDL) approaches to using the FSA administrative data within NASS are being explored. This report also contains information that may be helpful to the other efforts with respect to coverage and timing issues when considering entire state datasets.

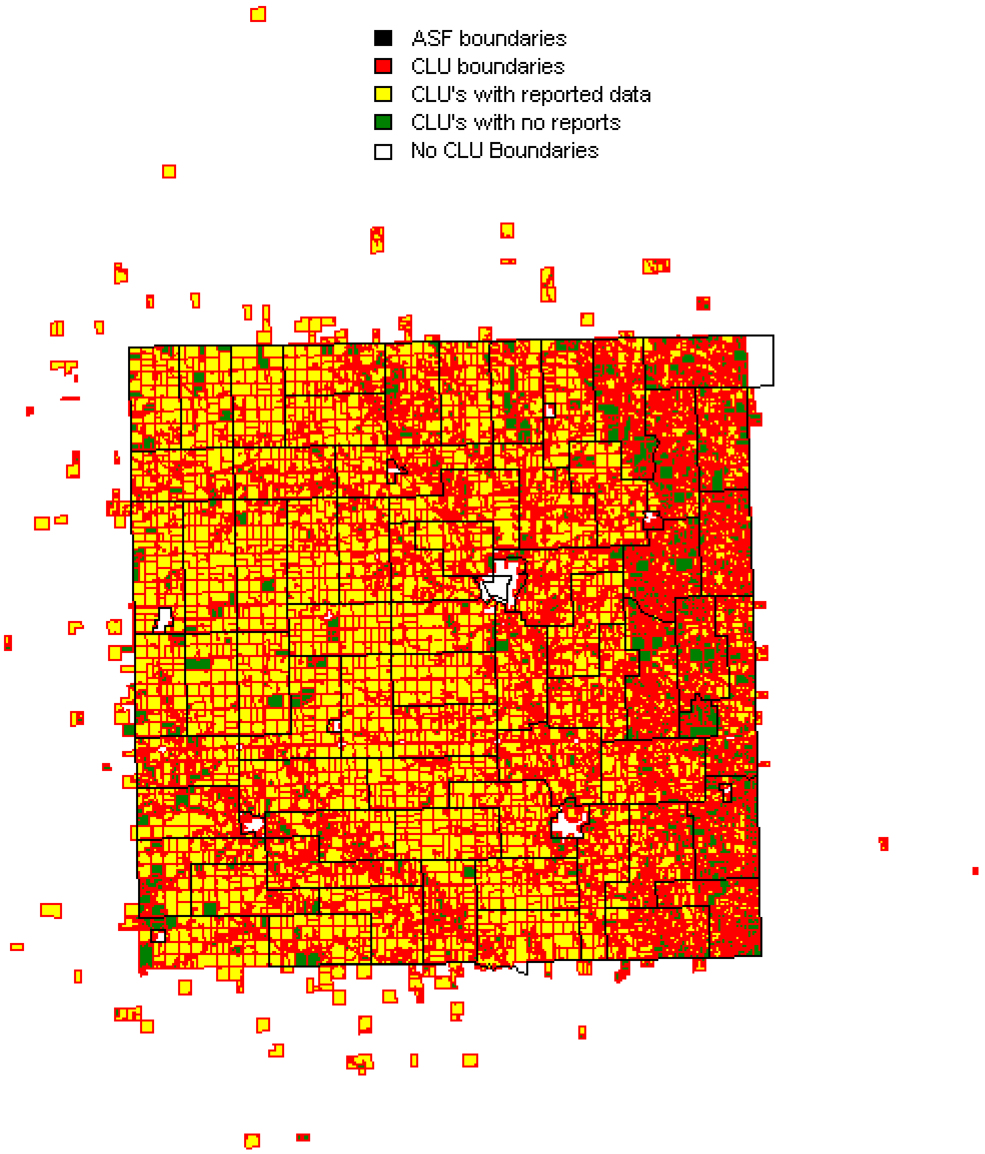
DESCRIPTION OF FSA DATA

Two types of administrative data were available from the FSA with respect to this project: digitized field boundaries for each county and one large state level file containing farmer signup reports. As of now, these two types of data are kept in two different computer systems and must be matched as needed. Tony Dorn has pulled the state level signup reports down to the NASS LAN for the 2004 crop season, and has provided a link to an FTP site created for storing the digitized field boundaries for all states [Dorn, 2005].

The FSA Common Land Unit (CLU) system creates digitized polygon boundaries of semi-permanent 'fields' in ERSI shape file format. There is one shape file per county. One problem with the county approach is that a farmer can report, in a single county office, for all crop land he operates in the state, irrespective of the county where the fields are physically located. An FSA field may or may not correspond to a field under the NASS area frame approach; usually it does not. NASS 'fields' contain one cover type, although some waste area is allowed. A CLU field is based on permanent boundaries and may contain one or more major cover types. The CLU boundaries were digitized by the Nebraska county FSA offices. CLU boundaries are digitized on PC screens with a 1-meter resolution digital image as the base layer. Some areas such as urban, water bodies and large non-agricultural zones are not digitized, and exist only as holes in the CLU shape file(s).

Figure 1 displays the digital FSA CLU shape file of Seward County, Nebraska for 2003. Primary sampling unit (PSU) boundaries from the NASS area sampling frame are shown in black to approximate the county's political boundaries. Note the 'halo' effect of the CLU polygons outside the political boundaries. These are areas reported by farmers at the Seward County office and administered by them, but fall outside the county boundaries. Areas shown in yellow have reported data for 2003, while areas shown in green have no reported data in the Seward County file. The green areas may have reported data which is administered in another county, or they may not have any reported data. White areas are 'holes' which are not digitized; there are additional holes for things such as roads not visible due to the scale of the Figure.

Figure 1. Digitized CLU Shape File for Seward County, Nebraska



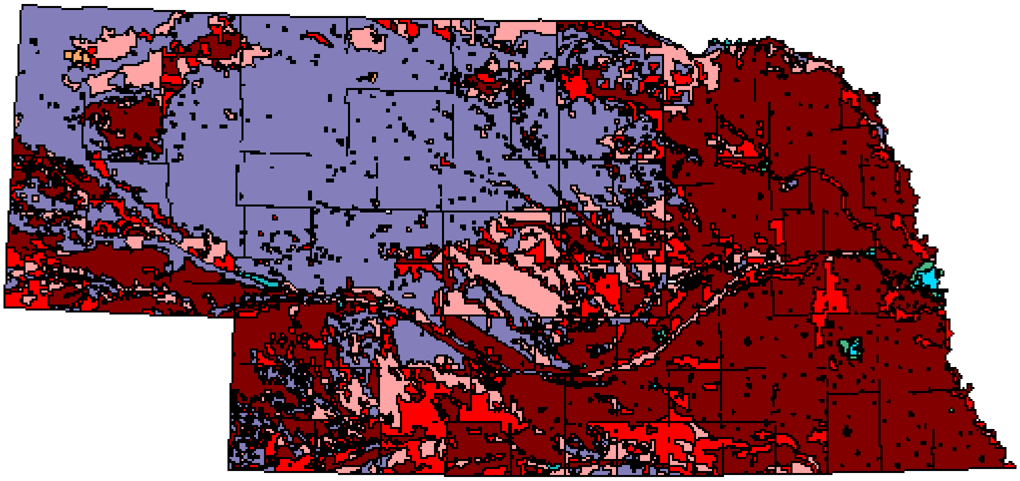
An FTP website exists for the purpose of downloading of county CLU shape files; however, at the present time, this site is at least one year behind for each state, if any data exists there at all. CLU files for this project were obtained directly from the Nebraska FSA state office. Although there are specific due dates for the certification of reported acres and the certification of county boundaries in each CLU, the FSA county offices continually update their respective CLU files as new data come in; thus there is never a final set of CLU shape files for a given state. Two sets of CLU files were delivered by the FSA for our use: the most current CLU as of October 2003 (used for 2002 and 2003 crop years) and, in August 2004, the most current available CLU shape files for the 2004 crop season.

FSA administrative program crop sign-up data, known as Form 578 data, are maintained in the FSA NITC mainframe at Kansas City, MO. These data are submitted to FSA headquarters and combined into a state-level file. Since the sign-up is voluntary, not every producer makes use of this program. The state level Form 578 files are available to NASS via a secure intranet login to the FSA HQ archive. Although there are specific due dates for reporting and certifying crop signup data, in practice the data are accepted throughout the year and the Form 578 files are subsequently updated. There is no 'final' data set for Form 578 data. There were three different 2003 sets of 578 data considered in this analysis: the dataset delivered to the NASS Crops Branch for use in the annual estimation program, the most current set as of October 2003, and a 'final' crop year 2003 set obtained from the FSA archive in June 2004. The crop data merged onto the 2003 CLU shape files is a subset of the October 2003 Form 578 dataset, consisting of those CLU polygons with ID values actually matching the entire state 578 dataset. This difference will be discussed in a later section.

NASS INPUTS AND PROCEDURES

Three types of input data were provided or obtained by NASS for use in this investigation: area frame segment (and internal field) boundaries for NASS area frame segments in agricultural strata, farmer reported June Survey data by field for the same area segments, and full scene Landsat digital imagery. At least in part as support for this research, Nebraska was added to the CDL Project in 2002. Analyses for this Project will focus on the three NASS area frame strata with the intensive, highest percent cultivation (Strata 11, 12, and 20) and would only show the remaining urban or non-agricultural strata (Strata 31 through 50) in summary discussions. Farmer reported data for all Nebraska area frame segments were captured from the June Area Survey edited data files by the NASS Spatial Analysis Research Section (SARS) staff. Digital Landsat imagery was obtained from the USDA Foreign Agricultural Service (FAS) Imagery Archive under a NASS/FAS cooperative agreement. Figure 2 depicts the NASS Nebraska area sampling frame; stratum 11, 12, and 20 are colored maroon, red, and pink respectively.

Figure 2. The Nebraska Area Sampling Frame



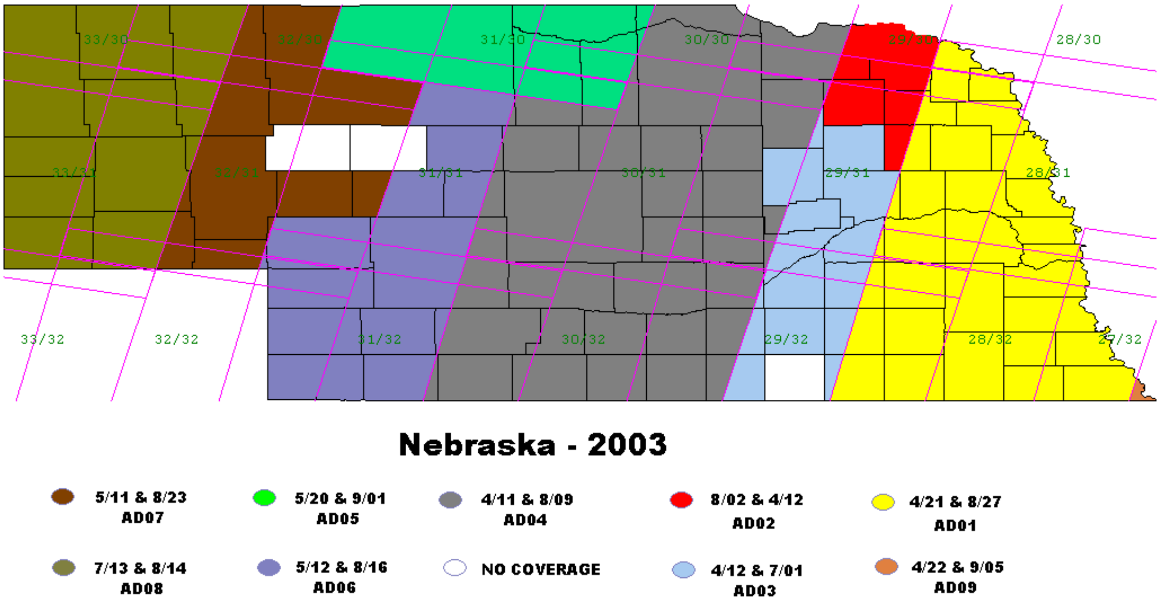
NASS field level segment data from the corresponding years' June Area Survey (JAS) were captured from the operational JAS 'D4' files and converted to formats usable for the CDL project. This is a normal part of the NASS CDL project each year. Nebraska field office staff, with significant help from the Geography Departments of the University of Nebraska – Lincoln and the University of Nebraska - Omaha, digitized the segment internal field boundaries of the operational NASS area frame segments based on the June area survey segment enumeration photos. The base layer for NASS digitization is 30-meter resolution Landsat imagery. Table 1 shows NASS stratum definitions, population numbers, and sample segment counts for the three years.

Table 1. The NASS Area Sampling Frame for Nebraska, 2002-2004

Stratum	Total Sq. Miles	Target Size	Population Segments	Samples 2002, 2004	Samples 2003	Stratum Description
11	30112	1.00	30202	285	240	>80% Cultivated
12	8755	1.00	8794	77	63	51-80% Cultivated
20	9531	2.00	4785	63	56	15-50% Cultivated
31	649	0.25	2610	4	4	Agri-Urban
32	168	0.10	1677	2	2	Dense Urban
40	27695	4.00	6915	40	32	<15% Cultivated
50	180	1.00	184	2	2	Non-Agricultural
State	77090		55167	473	399	

It takes all or part of sixteen Landsat V Thematic Mapper (TM, launched in 1984) or Landsat VII Enhanced Thematic Mapper (ETM, launched in 1999) scene path/row locations to completely cover Nebraska. Wherever possible, NASS will use two dates of imagery at each scene location. The preferred date combination for crop discrimination in an area such as Nebraska is one spring scene and one mid-summer scene. Cloud cover determines the extent to which the preferred date combination is achieved. Figure 3 displays the Landsat coverage dates used in the operational NASS CDL program for Nebraska, 2003. Analysis Districts (AD01-AD09) were determined by combining areas with the same date combination. Note that AD09, a small area in Southeastern Richardson County, was actually obtained from the Iowa 2003 analysis, and will not be covered here.

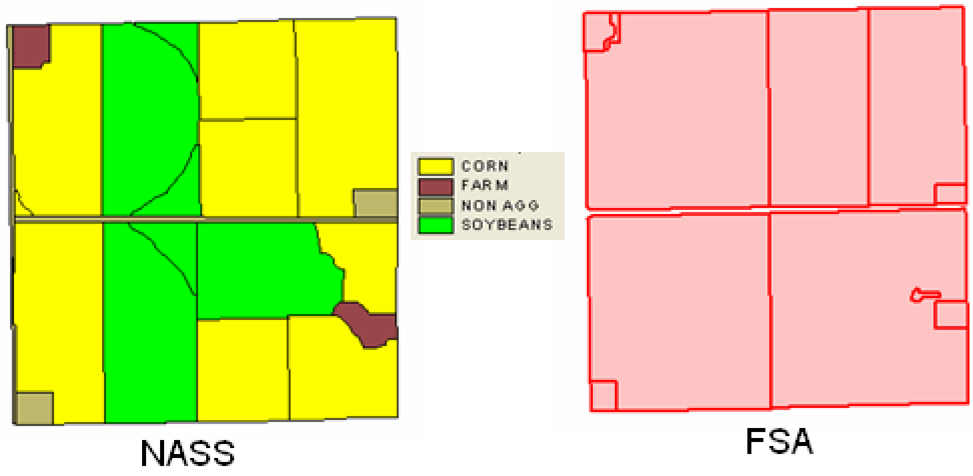
Figure 3. Landsat Scene Locations & Coverage Dates – NE03



AREA COMPARISONS – SEGMENT LEVEL

NASS SARS and Area Frame Section (AFS) staff used the outer boundaries of NASS segments to subset or cut out the corresponding CLU data from the FSA county shape files for the 2003 and 2004 seasons. The subset data files had to be manually edited where the CLU boundaries were split by the NASS outer segment boundary. This was used to generate a set of 'FSA segments' matching the area of NASS segments, but with internal boundaries as determined by the CLU polygons. Figure 4 displays a NASS segment beside the matching FSA CLU polygons (with its multiple crops per CLU).

Figure 4. NASS Segment with Matching CLU Boundaries



For the 2003 season, each CLU record contained information on total CLU acres (field name CALCACRES, whether or not the CLU had any reported data), plus acreage and cover information on up to seven reported cover types (field names RPTACRE1-7,CROP1-7). Each (RPTACREn,CROPn) combination described represents the sum of one cover type, although there may have been more than one report of that cover type in the CLU. This data was re-summarized by seven cover types within each record: corn, soybeans, wheat, sorghum, fallow (idle cropland), Conservation Reserve Program (CRP), and one category of all other ('rest'). Table 2 displays the re-summarized FSA data versus the NASS June Area Survey segment data, by NASS area frame strata.

Table 2. NASS JAS Segments Versus FSA CLU Subsets (2003)

	Average Acres per Segment					
	Stratum 11		Stratum 12		Stratum 20	
Cover	NASS	FSA	NASS	FSA	NASS	FSA
CALCACRES	645	627	655	621	1271	1277
Reported (Sum)	645	399	655	351	1271	362
No Report (Sum)	0	230	0	339	0	1003
Corn	215	165	143	102	95	59
Soybean	137	102	75	54	43	22
Wheat	39	26	35	23	48	39
Sorghum	13	10	11	13	12	8
Fallow (idle crop)	29	20	14	10	18	22
CRP	23	16	32	15	22	30
Rest/Other	189	60	345	135	1033	182

Total acres differences between the NASS segments and the FSA subset can be attributed to the fact that FSA CLU polygons sometimes do not include roads; FSA boundaries are digitized on both sides of major roads while NASS total acres represent all acres within the outer boundary. Comparing the Reported (Sum) acres in terms of the ratio between FSA and NASS, shows the total FSA reported data to be 61.8%, 53.6% and 28.5% of NASS total segment data, respectively by strata. Specifically for corn the ratios are 76.7%, 71.3%, and 62.1%; for soybeans they are 74.5%, 72.0%, and 51.2%. Missing data in FSA CLU polygons can be attributed to two sources: areas with no reports and areas where the reported Form 578 identification data did not match an existing CLU. Areas with no reports can also be broken down further into farmers who did not report and non-farm areas. At the segment level, it is impossible to determine which factor caused the differences between the NASS and FSA measures.

AREA COMPARISONS – TOTAL ACREAGE, ASD AND STATE LEVEL

Another approach to determining the coverage available from FSA CLU data is to compare total acreage measures. FSA CLU polygons do not include water bodies or urban areas, and sometimes also exclude pure non-Agricultural areas (refer to Figure 1). One objective measure of total area for any given state is the set of county polygons contained in the ESRI ArcView standard data set. This data set has no holes, such as lakes, large rivers or cities in it for Nebraska. Another measure of total area is the NASS Area Sampling Frame (ASF), which does have specific strata for major water bodies and different types of urban areas. For this analysis, an attempt was made to make the NASS ASF and the ESRI polygons comparable to FSA CLU polygons by subtracting the measured ASF acres for water and urban areas from both. Tables 3 and 4 show these measures of total acreage at the Agricultural Statistics District (ASD) and entire state levels.

Table 3. Measures of Total Acres

	FSA 2003	FSA 2003	NASS ASF	ESRI Area
ASD	Reported	CLU Sum	Comparable	Comparable
10	2651840	9706756	9084767	9092179
20	1136114	12507414	11969826	11983587
30	2493459	4841371	4840413	4820192
50	1953816	5278655	4677198	4693907
60	3422702	5200758	5168580	5133356
70	2164095	6551321	6018510	6015067
80	1624584	3038885	2871523	2861500
90	2461157	4507810	4240284	4286479
State	17907767	51632969	48871101	48886267

Table 4. Total Acres as Percent of ESRI Area

	FSA 2003	FSA 2003	NASS ASF
ASD	Reported	CLU Sum	Comparable
10	29	107	100
20	9	104	100
30	52	100	100
50	42	112	100
60	67	101	101
70	36	109	100
80	57	106	100
90	57	105	99
State	37	106	100

No attempt was made to characterize or compare the non-agricultural areas, which are a major portion of the total unreported acres. As would be expected, the percent reported to total area coverage is higher in more agriculturally intense areas. Also, the deletion of all urban ASF strata probably overcompensated for the urban areas not digitized by FSA.

Another hidden problem is that crop fields administered in another county office show up as fields with no data in the county where they physically reside. The CLU polygons, when not considering whether there are reported data in the CLU files, seem to cover the agricultural area well.

AREA COMPARISONS – MAJOR CROP ACREAGE, ASD AND STATE LEVEL

In order to evaluate the differences in major cropland areas, FSA CLU sum and Form 578 reported data are compared to NASS estimates for corn and soybeans. This comparison was made for three crop seasons, 2002-2004. As stated in a previous section, there is no ‘final’ data set for Form 578 data. The source of FSA reported data for each year is described below.

For 2002 and 2003 there are three different sets of 578 data considered in this analysis: one dataset as delivered to the NASS Crops Branch for use in the annual estimation program, the most current dataset as of October 2003, and a ‘final’ crop year 2003 dataset obtained from the FSA archive in June 2004. The crop data merged onto the two years CLU shape files are a subset of the October 2003 Form 578 datasets; consisting of those CLU polygons with ID values actually matching the corresponding year entire state 578 dataset. The 2004 data consist of the annual information delivered to the NASS Crops Branch plus the set of Form 578 most current as of early 2005 (subset from a SAS file with data for all states). The CLU Sum of Reported data for 2004 reflects merged data from the early 2005 file. Tables 5, 6, and 7 show major crop FSA reported data expressed as a percentage of NASS ASB estimates for the three years.

For 2002-2003, the prevalence of non-matches between the CLU polygons and the Form 578 was a major problem with respect to coverage. Approximately 24-26 percent of the data was missing at the state level for the major crops. However, the 2004 crop season results show only 5 percent differences when measured against the ASB final numbers. It is expected that this represents a maturing of the FSA process and will continue or improve in the future. A change in FSA farm programs could increase rather than decrease this difference in the future. Figure 5 reflects the upward coverage trend for the sum of CLU data (i.e. with Form 578 matches) at the state level.

Table 5. Major Crops as a Percent of ASB Final – 2002 Crop Season

	--------------------------Corn---------------------------				----------------------Soybean------------------------			
	FSA 578	FSA 578	CLU Sum	FSA 578	FSA 578	FSA 578	CLU Sum	FSA 578
ASD	Crops Br	Oct 2003	Oct 2003	June 2004	Crops Br	Oct 2003	Oct 2003	June 2004
10	98	99	62	99	96	96	39	96
20	99	106	57	106	99	99	53	99
30	99	100	69	100	100	100	70	100
50	99	100	83	100	100	100	85	100
60	99	100	85	100	99	99	84	100
70	98	102	54	101	99	100	41	99
80	99	100	84	100	100	100	84	100
90	99	99	71	100	100	100	73	100
State	99	100	74	100	100	100	76	100

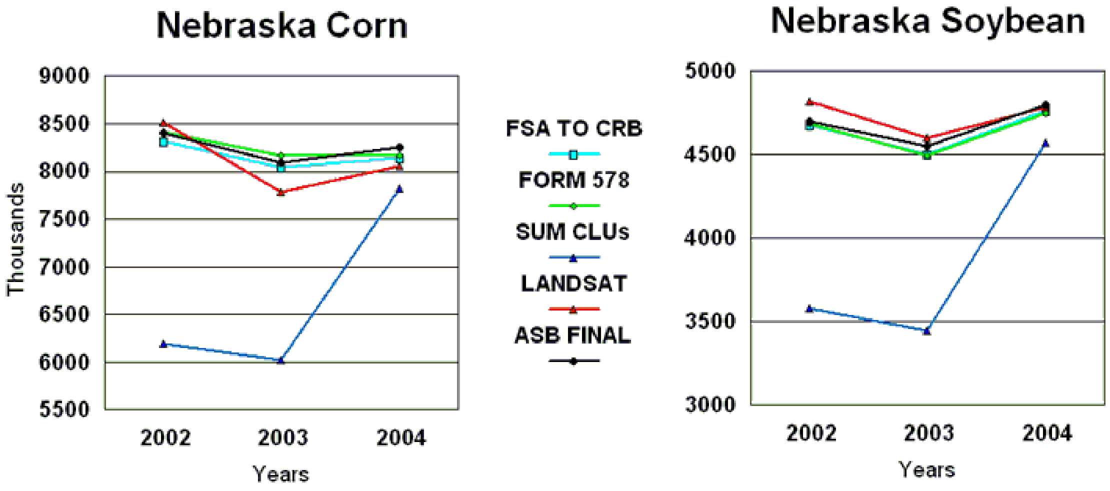
Table 6. Major Crops as a Percent of ASB Final - 2003 Crop Season

	--------------------------Corn---------------------------				----------------------Soybean------------------------			
	FSA 578	FSA 578	CLU Sum	FSA 578	FSA 578	FSA 578	CLU Sum	FSA 578
ASD	Crops Br	Oct 2003	Oct 2003	June 2004	Crops Br	Oct 2003	Oct 2003	June 2004
10	99	102	66	99	109	343	82	109
20	98	105	55	104	99	101	54	99
30	99	100	68	100	99	98	69	99
50	99	100	76	100	99	99	77	99
60	99	100	85	100	99	99	83	99
70	99	104	60	103	99	102	49	98
80	99	101	88	101	99	99	87	99
90	99	100	72	100	99	99	73	99
State	99	101	74	101	99	99	76	99

Table 7. Major Crops as a Percent of ASB Final - 2004 Crop Season

	------------------Corn-------------------			----------------Soybean---------------		
	FSA 578	FSA 578	CLU Sum	FSA 578	FSA 578	CLU Sum
ASD	Crops Br	Mar 2005	Mar 2005	Crops Br	Mar 2005	Mar 2005
10	99	99	94	93	93	90
20	99	99	93	99	97	93
30	99	99	94	99	99	95
50	99	100	94	99	99	94
60	99	99	96	99	99	96
70	99	99	94	99	97	92
80	99	99	95	99	98	95
90	99	99	96	99	99	97
State	99	99	95	99	99	95

Figure 5. FSA State Estimates of Major Crops



USING FSA CLU POLYGONS FOR NASS CDL SIGNATURE DEVELOPMENT

Several differences between NASS and FSA data were seen after reviewing the FSA dataset. The most important difference is between the definitions of a NASS ♧field♣ and a FSA CLU. In NASS segments, a field is defined as an area with one contiguous cover type. A NASS field can have a certain amount of ♧waste♣, usually considered to be no more than 5-10 percent, and still be usable for remote sensing ground ♧truth♣ training. Under the FSA system, multiple crops or cover types can be reported within a given CLU boundary during a specific crop year. The Nebraska 2003 database as delivered allows up to seven cover types in a CLU, while the Wisconsin Form 578 data has double digit numbers of cover types for some CLU polygons. A polygon containing more than one cover type cannot be easily used for ground truth training.

Comparing the images for the segment shown previously in Figure 4, we see that three fourths of the FSA data would not be usable for training; only the two corn fields in the upper right quarter would meet the criteria of one main cover type accounting for at least 90% of the CLU area.

For the 2003 dataset, the cover types in each CLU were sorted by size and the largest designated as CROP1. Then an automated edit defined a CLU as ♧usable♣ for training when the sum of acres for cover types listed in the CROP2-CROP7 fields is less than 10 percent of the CROP1 area. Table 8 compares the pixels available for training for NASS and FSA ground truth approaches in the NASS segment areas.

Table 8. Number of Segment Pixels, Total and Usable for Training Entire State, Nebraska 2003

	NASS		FSA		FSA as % of NASS	
Cover	Total	Training	Total	Training	Total	Training
Alfalfa	85152	22546	39558	19024	46.46	84.38
Beets	1283	647	1464	1044	114.11	161.36
Buildings	79	0	0	0	0.00	n/a
Corn	411959	145973	325597	142568	79.04	97.67
Crop Past	5968	2008	0	0	0.00	0.00
CRP	0	0	38094	26440	n/a	n/a
Dry Beans	5655	1709	3675	910	64.99	53.25
Fallow	50693	15242	47898	11059	94.48	72.56
Farm	13808	805	0	0	0.00	0.00
Grass	0	0	49349	21431	n/a	n/a
Hay	2240	482	0	0	0.00	0.00
Idle Cropland	52498	21241	0	0	0.00	0.00
Millet	3022	1652	6706	949	221.91	57.45
Non Agric	149796	15817	0	0	0.00	0.00
Oats	3370	633	3850	389	114.24	61.45
Other Crops	11152	4926	1737	607	15.58	12.32
Other Hay	14432	1537	0	0	0.00	0.00
Perm Past	459907	227463	3205	2281	0.70	1.00
Popcorn	564	302	0	0	0.00	0.00
Potatoes	618	370	585	518	94.66	140.00
Roads & RR	538	0	0	0	0.00	n/a
Rye	3360	472	102	15	3.04	3.18
Sorghum	24010	4742	22729	8069	94.66	170.16
Soybeans	243137	78446	181077	77331	74.48	98.58
Sunflowers	753	377	4938	0	655.78	0.00
Unknown	87	0	810060	0	n/a	n/a
Urban	30254	14463	30254	14463	100.00	100.00
Waste	282	0	0	0	0.00	n/a
Water	45968	24199	45923	24199	99.90	100.00
Wild Hay	34693	11341	0	0	0.00	0.00
Win Wheat	84104	24170	55652	14135	66.17	58.48
Wood. Pasture	3376	825	0	0	0.00	0.00
Woods	35544	13976	26742	13889	75.24	99.38
Total	1782041	636364	1699735	379321	95.38	59.61

It is evident from the many ‘zeros’ in Table 8 that some discrepancies exist between cover types and definitions. For example, permanent pasture at NASS is defined as anything livestock can access or ‘walk through’, while it has a very specific definition at FSA. The non-agricultural categories are not represented at all in FSA polygons, this includes such cover types as: buildings, farmstead, ‘non-ag’ tracts, roads, railroads, and waste in crop fields. The NASS definitions include several types of hay crops not represented in the FSA cover types; although the FSA ‘Grass’ type may cover some of these pixels plus some of what NASS calls permanent pasture. Pixels for woods, water, and urban come from analyst selected training sites outside of the NASS segments and are approximately equal between the two approaches. However, there are some woods pixels inside NASS segments. For comparison purposes ten groups of covers were created as seen in Table 9.

Table 9. Groupings of Crop and Cover Types

GROUP:	Crops & Cover Types:
CLOUDSFIL	clouds, filler
CORNALL	corn, popcorn
IDLEFALW	idle crop, fallow
NONCROP	farm, non-ag, perm past, grass
OTHRCROPS	all other crops and hays
SOYBEANS	soybean
URBAN	Urban
WATER	Water
WINWHEAT	winter wheat
WOODSTREE	woods, wooded pasture, trees

The second difference between the NASS and FSA approaches, with respect to crops, concerns missing data. In a NASS area segment, there are no fields without a cover type label; a NASS enumerator will observe the fields whenever possible even if the farmer refuses to report data on the survey. With respect to FSA CLU polygons, there are two sources of missing data: first, there may be no current year reports (signups) from farmers and second, there may be no CLU polygon (mainly in non-agricultural areas, such as around and including cities and towns). A measure of the missing data was shown in Table 2, and is also reflected in the Table 8 columns on ‘FSA as % of NASS’. Table 10 breaks this down by Landsat Analysis District.

Table 10. Analysis District Pixels for NASS Segments Versus FSA Polygons

Analysis District	Usable for Training		Entire Segment		Missing
	NASS	FSA	NASS	FSA	FSA
AD01	76,180	80,934	290,644	273,822	104,868
AD02	67,552	61,742	224,453	210,706	76,631
AD03	125,631	95,681	375,764	357,444	132,473
AD04	147,637	71,461	389,925	369,747	216,399
AD05	26,109	7,254	48,131	47,915	35,931
AD06	106,755	32,526	239,431	232,038	137,269
AD07	46,406	13,989	100,157	97,200	52,402
AD08	40,094	16,217	113,536	109,747	52,784

Three sets of classification statistics were calculated to evaluate using FSA data as a source of training pixels. First was the NASS operational approach which uses NASS enumerator defined internal segment boundaries to develop a set of signatures. Next, the FSA CLU polygons, as clipped for NASS segment areas, was used to develop the FSA set of signatures and was measured against FSA boundaries. Finally, the classification derived from FSA signatures was measured against the NASS internal segment boundaries (assuming the NASS boundaries are the 'truth'). The Kappa statistic [Congleton, 1999] was used for comparison of percent correctly classified by the three approaches. Table 11 compares the Kappa values for the three major crops plus overall, and lists number of signatures derived versus the pixels available for training from both sources. Note that the overall category in the table reflects the ten groups of cover types described above in Table 9.

Table 11. Kappa Percent Correctly Classified, Number of Signatures and Pixels Used
Based on Alternative Signature and Boundary Sources, Nebraska 2003

	Signatures =	NASS	FSA	NASS	FSA	NASS	FSA	FSA
Area	Boundary =	NASS	FSA	NASS	FSA	NASS	FSA	NASS
	Cover	#Sign.	#Sign.	#Pixels	#Pixels	%Kappa	%Kappa	%Kappa
AD01	CORNALL	14	12	23404	28663	89.08	81.53	80.46
	SOYBEANS	10	10	19004	25156	83.33	78.34	78.12
	WIN WHEAT	5	8	1328	1562	94.94	95.29	67.07
	OVERALL	89	72	76180	80934	86.16	83.68	69.12
AD02	CORNALL	11	18	22227	26977	90.22	81.60	86.23
	SOYBEANS	9	16	14291	15827	91.44	88.34	87.82
	OVERALL	87	78	67552	62140	77.43	87.56	75.90
AD03	CORNALL	90	14	46552	45112	94.56	79.61	81.37
	SOYBEANS	40	12	31131	27088	92.08	78.39	77.86
	WIN WHEAT	9	2	911	2027	95.36	92.79	94.12
	OVERALL	231	63	125631	95681	88.73	80.17	66.41
AD04	CORNALL	9	11	29042	28016	91.46	83.03	90.99
	SOYBEANS	6	4	11632	8865	93.51	88.08	88.31
	WIN WHEAT	22	3	4248	3600	97.60	93.07	74.02
	OVERALL	168	95	147637	71461	87.15	86.14	68.55
AD05	CORNALL	3	1	1083	1001	100.00	97.17	100.00
	OVERALL	47	27	26109	6373	76.49	92.06	59.97
AD06	CORNALL	7	9	21109	10379	96.47	79.07	80.77
	SOYBEANS	2	1	2388	395	99.70	88.69	88.30
	WIN WHEAT	45	8	7276	5082	94.35	78.14	76.16
	OVERALL	188	73	106755	32526	93.38	85.53	58.40
AD07	CORNALL	9	9	1426	703	97.83	98.95	37.52
	WIN WHEAT	5	6	5734	918	96.42	98.13	83.46
	OVERALL	102	72	46406	13989	91.46	95.67	67.54
AD08	CORNALL	4	4	1432	1717	96.38	83.16	68.42
	WIN WHEAT	24	4	4471	946	93.90	66.36	29.48
	OVERALL	129	76	40094	16217	93.48	87.55	41.55

Using the NASS training data alone for signature development always led to more signatures per analysis district. The NASS only number of signatures generated ranged from 47 to 231, while using FSA training data only led to a range of 27 to 95. This is attributable to the discrepancy in the number of training pixels available for each approach; there are approximately twice as many in the NASS training set overall. Reviewing the Kappa percent correct statistics by cover type, the NASS Kappa values were larger in 16 of 19 cases. Two of the three cases where the FSA Kappa's were higher are characterized by a comparatively small number of FSA pixels available for training for that commodity; the third value is only marginally higher (0.36%). Comparing the Kappa overall values for NASS only data versus FSA only, the NASS numbers are larger in 5 of 8 analysis districts. Considering NASS only data versus FSA data with NASS boundaries, the NASS only data values are always larger, sometimes significantly so. Figures 6 and 7 show the statewide classifications for both approaches; yellow is corn, light green is soybean, purple is other small grains and hay, red is sorghum, dark green is woods, light orange is non-agricultural, and tan is fallow/idle .

Figure 6. Nebraska 2003 Classification – NASS Training Data

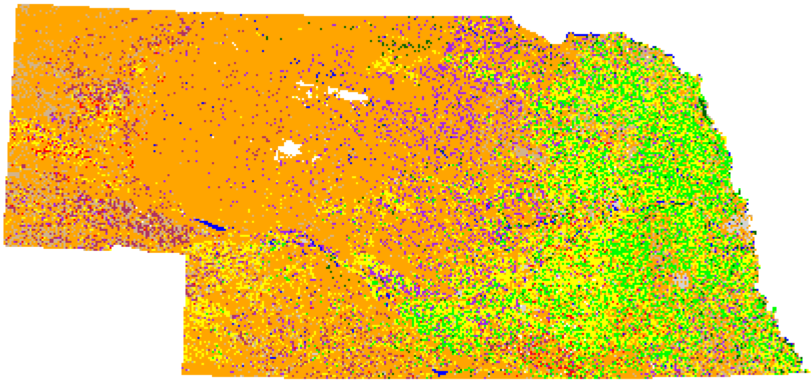
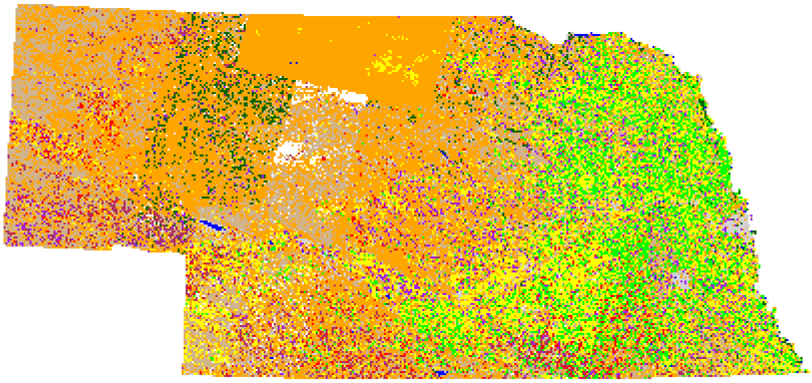
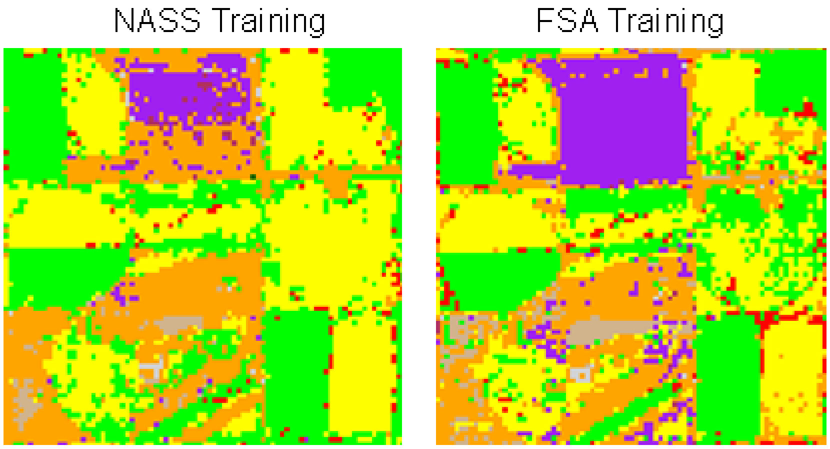


Figure 7. Nebraska 2003 Classification – FSA Training Data



Comparing Figures 6 and 7 at the state level, visual differences mainly exist in the non-agricultural categories (woods versus other non-agriculture) or in the distinction between grasslands (idle and fallow) with small grains and hay. However, a closer look, as seen in Figure 8, shows more speckle, or random misclassification, in the FSA image output.

Figure 8. NASS Segments vs. FSA CLUs as a Source of Training Data



Another measure of each classification is its performance in major crop estimation. The NASS operational approach uses categorized Landsat data as the auxiliary variable in a regression estimation procedure [Day, 2002]. In this regression, the dependent data (or y-variable) are the June Area Survey reported acres at the segment level for the specific cover type. The independent data (or x-variable) would be the number of pixels categorized to the cover type inside the outer segment boundary. Note that the two variables have different units, so the regression slope must also account for the conversion of pixels to acres (one pixel = 0.2222 acres). Thus a perfect relationship between the farmer reported data and the number of classified pixels would have a slope of 0.2222, and a r-squared coefficient of 1.000. The closer an individual regression is to these values, the 'better' fit it has. For comparison purposes, area frame strata 11, 12, and 20 were combined, although in practice they would be estimated separately whenever possible. Analysis District AD05 did not have enough segments to create a regression estimator. Tables 12 and 13 display the regression parameters and results using both sets of signatures versus NASS outer segment boundaries and farmer reported data.

Table 12. Combined Strata Regression Statistics – R-Squared
Nebraska 2003, NASS Boundaries

Analysis	Corn		Soybean		Winter Wheat	
District	NASS	FSA	NASS	FSA	NASS	FSA
AD01	0.780	0.568	0.803	0.707	0.891	0.565
AD02	0.870	0.774	0.844	0.879	0.984	n/a
AD03	0.922	0.693	0.931	0.785	0.980	0.850
AD04	0.926	0.873	0.947	0.912	0.949	0.829
AD05	n/a	n/a	n/a	n/a	n/a	n/a
AD06	0.964	0.782	0.987	0.876	0.844	0.748
AD07	0.928	0.422	n/a	n/a	0.938	0.845
AD08	0.957	0.776	n/a	n/a	0.863	0.499

Table 13. Combined Strata Regression Statistics - Slope

Nebraska 2003, NASS Boundaries

Analysis	Corn		Soybean		Winter Wheat	
District	NASS	FSA	NASS	FSA	NASS	FSA
AD01	0.2115	0.1727	0.2141	0.2010	0.2442	0.1878
AD02	0.2306	0.1973	0.2208	0.2104	0.2515	n/a
AD03	0.2253	0.1850	0.2391	0.2188	0.2750	0.1792
AD04	0.2457	0.2170	0.2290	0.2340	0.2664	0.2561
AD05	n/a	n/a	n/a	n/a	n/a	n/a
AD06	0.2355	0.1806	0.2499	0.2315	0.2682	0.2308
AD07	0.2485	0.3505	n/a	n/a	0.2567	0.3051
AD08	0.2062	0.2291	n/a	n/a	0.2209	0.1992

CONCLUSIONS AND RECOMMENDATIONS

This section is arranged into two parts, a textual description of the information learned when attempting to answer the five questions posed in the Introduction, and a numbered list of the specific problems and my recommendations as to solving them.

The first question to answer was how well the FSA and NASS data sets compare over the same areas. Considering the 399 segments from the 2003 June Area Frame Survey, and comparing the Reported (Sum) acres in terms of the ratio between FSA (matched CLU/578) data and NASS data, we see the total FSA reported data to be 61.8%, 53.6% and 28.5% of NASS total segment data, respectively for strata 11, 12, and 20. However, the stratum ratios improve for major crop types. Specifically for corn the ratios are 76.7%, 71.3%, and 62.1%; for soybeans they are 74.5%, 72.0%, and 51.2%. In an attempt to consider total area measures, the FSA CLU reported data (merged from 578 files), the sum of FSA CLU areas regardless of reported data status, and the NASS Area Sampling Frame Nebraska (with cities and water removed) were compared in size to that of the political boundaries found in the ESRI standard dataset for the state of Nebraska (also with cities and water removed). Under this measure, FSA reported data is 37% of the ESRI ‘comparable’ standard, the sum of CLU polygons is 106%, and the NASS ASF is 100%. The sum of CLU polygons being 6 percent higher probably reflects cropland areas in the NASS ASF ‘ag-urban’ stratum 31, which were removed from the ASF and from the ESRI data and the fields that are administered in another county office. Tables 5, 6, and 7 depict that differences between the reported FSA data versus the NASS and ESRI measures are highly variable with respect to Agricultural Statistics Districts which are contiguous groups of counties. Again we see this difference tempered significantly when considering only corn and soybeans, with the NASS Agricultural Statistics Board (ASB) official estimates as the standard. FSA reported acreage for corn and soybeans was in the range of 74-76% of the ASB final for 2002-2003, but rose to 95% for the final 2004 FSA number. Overtime, as the FSA county offices become more comfortable with the systems involved, one would expect the percent of matches between CLU and 578 data to increase. However, if FSA programs change, such as with a decrease in payment levels, this difference might increase. There will always be some percentage uncovered (not reported) or un-matched however, leading to the conclusion that as a sampling frame, the FSA CLU/578 match is incomplete. Nebraska was one of the flagship states for the implementation of the CLU approach, and one would expect each state to have its own adjustment problems over their first years in operation.

The next question to be addressed is whether the FSA data alone can be used to generate spectral signatures for classification of the CDL. There are some definitional differences between FSA reported data and NASS field labels that must be worked out. For example, all hay crops (other than alfalfa) are considered 'Grass' to FSA; plus the definitions of pasture and permanent pasture are intertwined in the two approaches. Groupings of cover type names between NASS and FSA alleviate this problem somewhat. Comparing the classifications using signatures made solely by FSA inputs versus the standard NASS approach, measures of the percent of known pixels correctly classified show the NASS values generally better for corn, soybeans, and winter wheat but the FSA values are not unreasonable. However, when measured as a predictor variable in a regression approach, the FSA values fall behind even more. R-squared values for NASS derived classifications range from 0.780 – 0.964 (corn), 0.803 – 0.987 (soybean), and 0.844 – 0.984 (winter wheat) versus those from FSA derived signatures of 0.422 – 0.873 (corn), 0.707 – 0.912 (soybean), and 0.499 – 0.850 (winter wheat). Under these measures, FSA should not be used alone for signature creation, unless accompanied by a major manual (remote sensing analyst) edit to determine the differences. This might also involve re-defining what is 'usable for training' from the FSA polygons, i.e. only large fields or only those fields that can be compared to current year imagery by an analyst (manual review). Finally, the addition of 'sub-field' boundaries which is currently being considered at FSA may remove a major portion of the variation currently introduced when using the FSA CLU polygons for training. We attempted to remove this variation by only using CLU boundaries for training if one major cover type accounted for 90 percent or more of the acreage in the CLU, but this approach may not be sufficient.

The next question is a corollary to the one above, if it should not be used as stand alone can it be used to augment existing NASS training data from the segments. The answer to this question would be similar to the one above; involving a re-defining of what is 'usable for training' from the FSA polygons, i.e. only large fields or only those fields that can be compared to current year imagery by an analyst (manual review). The positive side of this particular answer is that the sample of polygons derived from FSA data can be specifically targeted to cover types, increasing data for some that may be insufficiently covered by NASS segment data. We have sufficient training data for the two major crops in Nebraska, but could use supplemental winter wheat and other crops information. When a state has a very thin training sample from the JAS even for its major crops, we should be able to use the FSA CLU polygons for additional training data. We are currently looking at this approach in Florida for classification of 2004 crop year imagery.

The next question specifically asks about using FSA data for additional training information for 'minor' crops. Unfortunately, this question is hard to answer when using Nebraska as the model. The area devoted to 'minor' crops in Nebraska is small in general, and although the classification of one or two (such as potatoes) might benefit from a targeted approach as previously described, overall it will not make much of an impact. However, in some other states this will not be the case. For example, we are already looking at using FSA polygons for minor crops in Wisconsin; these make up a larger percentage of the planted cropland than in Nebraska. We must work carefully with FSA for minor crops, because the merge of Form 578 data onto CLU polygons does not seem to be standardized by state yet, and minor crops are the most likely to have problems.

Finally, how do we best convert FSA data into a form usable in our CDL system? There are several parts to this answer. First, we must ignore the concept of a segment when considering FSA polygons and use them as a population of individual training areas. Then we can target crops and cover types which need more information and go after those alone. We then need an automated way to turn a selected CLU polygon into a NASS training site recognizable by the PEDITOR system. This involves converting a shape file boundary into a PEDITOR labeled mask; current PEDITOR modules only convert shape files into strata boundary files (not segments).

The following bullet points address the problems and recommendations directly:

Problem 1: Dataset is incomplete

- Either no signups at all or Form 578 signups not matching CLU ID's
- Caution: failure to signup may not be limited to small farmers
- CLU Non-matches with Form 578 significant 2002, 2003
    - 2004 closer, less missing matches per CLU/578, hopefully a trend
- Never Static – both CLU and Form 578 can be updated at any time

Recommendation(s) For Problem 1:

- Will always have to 'freeze' the FSA datasets at some point and work from there
- FSA is working toward creating one system with all data (no separate '578' file)

Problem 2: Multiple covers per CLU polygon

- Seven possible per CLU in NE 2003
- Pre-merge done by FSA lost some info such as which type corn or wheat
- Double digit number of covers per CLU seen in Wisconsin 2004

Recommendation(s) For Problem 2:

- This will remain a problem for use as remote sensing training
- We will have use the subset of CLU polygons containing one cover (90% +)
- Possible FSA 'sub-field' digitization in the future will make more fields usable

Problem 3: Data administered by one county office that is 'politically' in another

- Unclear how this is handled in the dataset
    - Probably added to the number of non-matches CLU/578 seen in Problem 1
- Possibly not standard between county offices

Recommendation(s) For Problem 3:

- Get better description from FSA as to how this is handled
- Hopefully, they will have a standard in place that we can use/detect

Problem 4: Double Cropping

- Unclear how this is handled in the dataset
- Only a minor amount in Nebraska, not enough to determine standards

Recommendation(s) For Problem 4:

- Get better description from FSA as to how this is handled

Problem 5: Shape Files and Form 578 datasets are hard to convert to PEDITOR

- Data for each of the three years came in different formats, from different sources
- Current approach (as used for this analysis) is manually intensive
- CLU Polygons generated to look like NASS segments

Recommendation(s) For Problem 5:

- Need a standard format of CLU and its crops database from FSA
    - Especially if Form 578 data is merged onto it
    - Get Sub-fields (one cover type) format, soon (?) coming from FSA
- Forget the ‘segment’ approach, and consider each polygon on its own merits
    - Is it usable for training in a remote sensing analysis?
- Editing will have to be done to determine those CLU ‘fields’ usable for training
    - Sampling of the usable fields to result in an ASCII list of ID’s
- Create a new PEDITOR module to sample fields from FSA county CLU files
    - Module CLUSEGS already under development
    - Subset a county CLU shape file based on an ASCII list of ID’s
    - Create associated PEDITOR format ‘segment’ files for each
        - If crop (578) data pre-merged, will create ground labels

REFERENCES

Congalton, R.G. and K. Green (1999). *Assessing the Accuracy of Remotely Sensed Data: Principles and Practices*, Boca Raton, FL: CRS Press, Inc.

Craig, M. (2001). The NASS Cropland Data Layer Program. Presented at the Third International Conference on Geospatial Information in Agriculture and Forestry, Denver, Colorado, 5-7 November 2001.

Craig, M. (2001) "A Resource Sharing Approach to Crop Identification and Estimation", ASPRS 2001 Annual Convention Technical Papers, St. Louis, Missouri, April, 2001.

Day, C. (2002), A Compilation of PEDITOR Estimation Formulas. RDD Research Paper RDD-02-03, USDA/NASS, National Agricultural Statistics Service, Washington, D. Jan. 2002.

Dorn, Tony (2005) *FSA Data Potential for NASS*, draft NASS white paper, May, 2005.

Hanuschak, G., Hale, R., Craig, M., Mueller, R., and G. Hart (2001), The New Economics of Remote Sensing for Agricultural Statistics in the United States. *Proceedings of the CAESAR Conference, June 2001*, Italian Statistics Agency (ISTAT), Rome, Italy

APPENDIX I
Joint FSA/NASS 2001 Research Project in Southeastern Nebraska

Introduction

Discussions on the joint research project were initiated in the Spring of 2001 to look for mutual benefits between the NASS Cropland Data Layer and FSA♣s Common Land Unit GIS. The final objectives of this project were set during a meeting between staff from the USDA/NASS Geospatial Information Branch (George Hanuschak, Bob Hale, and Mike Craig) and the USDA/FSA Production Estimates and Crop Assessment Division (Glenn Bethel) on August 30, 2001. The objectives as stated then were:

1. Categorize the Landsat data for the five county area using NASS's standard procedures and only the NASS June area segment data for ground data.

2. Categorize the same area using FSA data as ground data, more minor crop breakouts if possible.

3. Compare NASS segment data with FSA data for NASS segment areas.

4. Compare NASS categorized data with FSA for the entire area and not just NASS sampled areas.

5. Look at methods to improve classification accuracy with additional FSA data as input (related to item 2).

Five counties (Gage, Jefferson, Lancaster, Seward, and Saline) in Southeastern Nebraska were chosen to be the focus of the Joint project. These counties together make up a rectangle that is wholly contained in just one Landsat scene and have enough NASS June area segments overall to make a reasonable classification without the added data available from the FSA administrative data system. The main summer crops of interest in this area are corn and soybean; although sorghum, winter wheat, and alfalfa are present here also. Although results for winter wheat are used in this analysis, wheat estimation would possibly involve a different set of imagery dates in future usage. Software for this Project includes: the NASS PEDITOR software system for image processing and some geographic information system (GIS) functions, ESRI♣s ArcView software was used for some GIS functions, and MicroSoft♣s FoxPro was sometimes used for conversion of ground data between the FSA and PEDITOR formats.

Input Data From FSA

Two types of data were provided from the FSA Common Land Unit (CLU) system: shape files containing CLU boundaries as digitized by Nebraska county FSA offices, and the 2001 final FSA crop administrative program sign-up data (Form 578) corresponding to the CLU boundaries. The shape files were obtained from the FSA county offices, and the 578 sign-up data from the central FSA system in Kansas City. FSA headquarters staff merged the two datasets to create one shapefile for each county with both the boundaries and sign-up data. The FSA HQ group also provided geo-referenced TIF files of Landsat imagery for the five counties to aid in the CLU display in ArcView.

Several differences between NASS and FSA data were seen after reviewing the FSA dataset. The most important difference is between the definitions of a NASS ♧field♣ and a FSA CLU. In NASS segments, a field is defined to be an area with one contiguous cover type. A NASS field can have a certain amount of ♧waste♣, usually considered to be no more than 10 percent, and still be usable for remote sensing ground ♧truth♣ training. Under the FSA system, multiple crops or cover types can be reported within a given CLU boundary (their database allows up to seven cover types in a CLU) during a specific crop year. A polygon containing more than one cover type cannot be easily used for ground truth training. For the purposes of this study, a CLU is defined as ♧usable♣ for training as being CROP1 when the sum of acres for cover types listed in the CROP2-CROP7 fields is less than 10 percent of the CROP1 area. Table 1 shows, for the entire 5-county area, the number of overall CLU♣s by cover type in the CROP1 field (usually the largest cover type present in a specific CLU) and the number of CLU♣s ♧usable♣ for remote sensing training.

Table 1: Number of CLU♣s by Crop Type in Merged FSA Five County Data Set

FSA CROP1 Description	# CROP1 CLU♣s > 0	Total CROP1 Acres	# CLU♣s Usable For Training	Acres Usable for Training
Alfalfa	1,061	16,362	805	9,409
Corn	6,999	254,190	4,893	145,098
CRP	3,138	47,217	3,028	44,674
Fallow	163	1,630	129	891
Forage / Soybean & Sorghum	58	540	48	393
Grass	3,501	77,177	2,085	35,600
Oats	64	787	45	427
Sorghum(Grain)	2,722	82,059	1,623	34,789
Soybean(Grain)	7,770	257,801	5,630	153,713
Wheat	1,094	30,422	712	15,517
*Other (Not in NASS)	116	892	98	613
Total	26,686	769,077	19,096	441,124

* Included: Clover, Home Garden, Millet, Mixed Forage, Plums, Rye, Sunflower, Sorghum Forage, Pulp Trees, Triticale, Turn Areas, Waterbank, and Wildlife Habitat.

The second difference between the NASS and FSA approaches, with respect to crops, concerns missing data. In a NASS area segment, there are no fields without a cover type label; a NASS enumerator will observe the fields whenever possible even if the farmer is a refusal with respect to reporting data on the survey. With respect to FSA CLU polygons, there are two sources of missing data: first, there may be no current year reports (signups) from farmers and second, there may be no CLU polygon (mainly in non-agricultural areas, such as around and including cities and towns). Table 2 shows several attempts to measure the amount of missing data, with respect to NASS procedures, found in the FSA five county data set. Two ratios, expressed as percentages, were calculated: ♣578 Rptd to Sum CLU♣ is the ratio of Form 578 data reported data to the sum of CLU polygon data and ♣CLU Sum to ASF-Urban♣ is the ratio of the ♣CLU Sum♣ to the ♣Area Frame♣ corrected for urban (i.e., minus the ♣Urban 31+32').

Table 2: Area Measures Used to Review Missing Data in Study Region
In Acres Unless Otherwise Noted to be in Percent (%)

County =	Gage	Jefferson	Lancaster	Saline	Seward	5-County
FSA Form 578 Reported	245,386	176,585	181,226	228,097	165,271	996,565
Sum of CLU Polygons	535,768	357,106	469,028	356,918	355,489	2,074,309
Ratio 578 Rptd to Sum CLU	45.8(%)	49.5(%)	38.6(%)	63.9(%)	46.5(%)	48.0(%)
ESRI County Polygon Set	550,360	368,318	541,895	368,681	368,467	2,197,721
Sum of Landsat Pixels in ESRI	551,200	367,303	538,138	369,233	370,070	2,195,944
Area Sampling Frame (ASF)	558,080	372,608	537,088	368,960	370,496	2,207,232
Urban Strata (31+32) in ASF	7,296	3,392	54,592	4,928	4,864	75,072
Ratio CLU Sum to ASF - Urban	97.2%	96.7(%)	97.2(%)	98.0(%)	97.2(%)	97.3(%)

Reviewing Tables 1 and 2, we see that the ratio of FSA ♣578♣ reported data for the current year to the total CLU acreage is approximately 48 percent overall, which is lower than expected. A large amount of the missing data might be attributable to non-crop areas and pasture, etc. If you remove the NASS defined urban areas, the total CLU acreage is approximately 97 percent of the area sampling frame coverage. The three percent difference there is probably just in the definition of ♣urban♣. Looking at the portion of FSA reported data that is theoretically usable for training a supervised remote sensing classifier, we see that more than half of the acreage is usable.

Input Data From NASS

Three types of input data were provided by NASS: two dates each of full scene Landsat 7 ETM data for two scene areas, area frame segment (and internal field) boundaries for selected NASS area frame segments, and farmer reported June Survey data by field for the same area segments. In discussions with the NASS Nebraska field office, it was decided to obtain segment data for a larger area than just the five counties originally specified. This change was made for two reasons: the Cropland Data Layer (CDL) product for a larger region would be a better base for deciding on future CDL products and county acreage estimates in Nebraska (i.e., the inclusion of the whole state in the operational 2002 or 2003 NASS CDL Project) and the resulting classification of the 5-county area would be a more stable base for comparisons.

Analyses for this Project would focus on the two area frame strata with the intensive, highest percent cultivation (Strata 11 and 12) and would ignore the remaining extensive (Stratum 20) and urban or non-agricultural strata (Strata 31 through 50). Nebraska field office staff digitized the segment / field boundaries for seventy-three (73) area segments, sixty-two (62) in Stratum 11 and eleven (11) in stratum 12. Farmer reported data for all the 2001 Nebraska area frame segments were captured from the June Area Survey edited data files by the NASS Spatial Analysis Research Section (SARS) staff. Additionally, SARS staff created a subset of the five county CLU data containing only the twenty-nine (29) NASS area segment locations but with internal CLU boundaries as ♧fields♣.

The larger region to be covered for the initial NASS classified product was defined to be all counties and parts of counties (20+ counties) in Nebraska that fell within the borders of two Landsat scene locations: Path 28, Rows 31 and 32. Landsat scenes used from these two locations would be chosen to have the same dates to allow for signature extension across both scenes. The five county area specified for the NASS/FSA joint project is contained wholly in the Path 28 Row 31 scene location. Eleven scene dates were available for each scene location, each with varying amounts of cloud cover. Considering both cloud cover and crop progress, the dates chosen were May 1st and August 13th, 2001. Multi-temporal (14-channel) scenes were created by SARS for both scene locations by overlaying these image dates.

Objective 1. Categorize the Landsat data for the five county area using NASS's standard procedures and only the NASS June area segment data for ground data.

The first analysis done was to create the best possible classification of the two scene area based on all seventy-three (73) NASS segments available in the two scene Landsat area (Path 28, Rows 31 & 32). Using all NASS area segments as ground truth, and standard CDL methodology and procedures, a classification based on 107 cover type signatures (labeled analysis AD01) was produced for the two Landsat scene area. The NASS PEDITOR software system was used to perform image processing, classifier evaluation, and regression analysis.

As seen in Table 3, both corn and soybean had percent correct (Producer Accuracy) values over 90%, with commission errors under 10% (User Accuracy = ♣1 - commission error♣ thus was also over 90%). Regression r-squares for both crops in the AD01 analysis were in the 0.9 range, even with no outlier analysis and deletion as is usually done in the CDL approach. This was a very good classification under normal CDL criteria.

Table 3: Analysis Statistics for AD01 - All 73 NASS Segments in Two Scene Area

Cover Type	Percent Correct	Commission Error	Kappa Statistic	R-square Stratum 11	R-square Stratum 12
Corn	90.81	07.61	86.21	0.920	0.908
Soybean	92.67	08.77	88.86	0.893	0.882
Sorghum	85.87	21.13	85.38	0.909	n/a
Alfalfa	90.36	20.97	90.27	0.804	0.642
Win Wheat	97.96	16.20	97.92	0.903	n/a
All Covers	86.50	n/a	82.04	n/a	n/a

For later comparisons to FSA only data, a CDL type analysis of the 5-county area was done using only the reduced set of 29 segments, but with the original NASS ground truth labels and boundaries. This analysis was designated ♣AD04'. A signature statistics file of 86 categories was created and used for classification. Table 4 shows the classifier evaluation and regression analysis for AD04. Stratum 11 is shown for all segments and for the outlier reduced set of segments. Stratum 12 was removed from the regression analysis in Table 4 because it only has two segments in the reduced 29 segment subset. Stratum 12 is represented in the percent correct.

Table 4: Analysis Statistics for AD04 - 29 NASS Segments Only

Cover Type	Percent Correct	Commission Error	Kappa Statistic	R-square Stratum 11 (All 29 segs)	R-square Stratum 11 Outlier Adj.
Corn	93.24	4.89	91.15	0.918	0.952
Soybean	96.73	6.21	95.28	0.927	0.927
Sorghum	87.23	10.02	86.09	0.890	0.939
Alfalfa	98.98	21.14	98.97	0.694	0.771
Win Wheat	99.81	1.32	99.80	0.861	0.861
Overall	89.43	n/a	86.99	n/a	n/a

Using Table 4 as a guide, we see that the 29 segment area is reasonably representative of the entire 73 segment set. Only the r-squares for alfalfa and winter wheat were reduced in a significant amount, even with the outlier adjustment applied.

Objective 2. Categorize the same area using FSA data as ground data, more minor cropbreakouts if possible.

The second analysis done, labeled ♣AD02♣, was designed to determine how well the FSA CLU data could be used as ground truth for classifier training in the normal NASS CDL approach. As stated earlier, there are 29 NASS June area segments in the 5-county area. For comparison with NASS only data, the AD02 analysis was performed using CLU data from these 29 segment areas. New ♣FSA segments♣ for ground truth training were created by intersecting the outer boundaries of the NASS segments with CLU shape files. The intersection of the two sets of boundaries was accomplished using the ESRI ArcView software; the conversion of the new shapefile boundaries to PEDITOR format mask files with field labels was accomplished using a special PEDITOR module designed for this purpose.

PEDITOR format ground truth files were created using Microsoft FoxPro from the database portion of the shapefile. This was a somewhat torturous process because of the missing data and multiple crops per CLU differences mentioned earlier. During the Foxpro process, fields with missing data or multiple crops were labeled as ♣bad for training♣. Unfortunately, the new ♣fields♣ with multiple crops types were labeled only by the crop name in the CROP1 field with any other acreage (CROP2-CROP7) put in the waste category. Thus the ground truth files created could be used for selecting training fields, but presented a problem for the regression analysis (i.e., crops in the CROP2-CROP7 variables were not represented). A signature statistics file of 86 categories was created and used for classification. Table 5 shows the classifier evaluation and regression analysis for AD02. The percent correct, commission error, and kappa statistics shown use only the ♣usable♣ pixels, and thus are comparable to those from AD01. With the problem mentioned above plus a much smaller sample than in AD01, the NASS outlier deletion approach was used to stabilize the AD02 analysis. Stratum 12 had only two segments in the reduced data set, and thus it was not used in the regression analysis.

Table 5: Analysis Statistics for AD02 - 29 FSA CLU-based Segment Areas Only

Cover Type	Percent Correct	Commission Error	Kappa Statistic	R-square Stratum 11 (All 29 segs)	R-square Stratum 11 Outlier Adj.
Corn	85.86	15.85	79.32	0.294	0.760
Soybean	86.84	12.19	79.28	0.830	0.913
Sorghum	76.82	10.99	75.62	0.449	0.851
Alfalfa	76.15	0.00	76.03	0.202	0.347
Win Wheat	97.80	2.77	97.70	0.900	0.908
All Covers	84.85	n/a	79.53	n/a	n/a

This is a reasonable, although slightly worse, classification of the same segment areas as the NASS based AD04 subset mentioned earlier, with the exception of a significant drop in the corn and alfalfa r-squares. The corn problem seemed to be attributable to 3 bad signature categories, one a corn category that overlaps significantly with soybean, and 2 categories (one grass and one idle cropland) which look like corn. My guess is that these problems could have been caught in a manual review of the FSA CLU♣s/fields before clustering. Alfalfa has similar problems.

Another way to look at using the FSA data in the CDL approach is to train the classifier with FSA data but do the statistical analysis using NASS segment boundaries and reported data. This approach is an attempt to remove the FSA missing data problem from comparison analyses. A look at Table 6 will show that the classifier performance using only the FSA data for training, as compared to the NASS boundary and reported data, is significantly degraded from either that of using the NASS data alone, or using the FSA data alone.

Table 6: Analysis Statistics for AD03 - FSA Classifier Applied to NASS Boundaries (29 seg.)

Cover Type	Percent Correct	Commission Error	Kappa Statistic	R-square Stratum 11 (All 29 segs)	R-square Stratum 11 Outlier Adj.
Corn	68.32	38.90	56.61	0.310	0.765
Soybean	80.17	21.4	71.52	0.817	0.901
Sorghum	57.39	45.68	53.21	0.451	0.896
Alfalfa	39.80	41.79	39.59	0.200	0.625
Win Wheat	67.49	16.47	66.79	0.908	0.867
All Covers	53.55	n/a	43.76	n/a	n/a

This difference is felt to come from two sources: inherent definition problems between CLU♣s and NASS fields causing some mislabeled signatures, and a better review for bad fields done in the original NASS analysis. The fact that the outlier adjusted R-squared values for major crops are reasonable in AD03 shows that the definition problem is not insurmountable and a better manual review of the FSA based training fields would probably straighten out the signatures also. The ♧All Covers♣ Percent Correct statistic is low mainly because of the difference between naming conventions for non-crop covers (such as Permanent Pasture versus Grass, etc.) and should be discounted.

The second part of Objective 2 was to consider creating more minor crop breakouts if possible. In the context of the CDL approach, this would mean measuring the ability of the FSU polygons to create classifier signatures for non-major crops or cover types. Table 7 shows the number of fields and acres of cover types found in the 5-county CLU data and not found (with the exception of Oats) in the 73 NASS segments. Oats is included in the Table 7 because only one field was found in the NASS segments, not enough to create a valid classifier signature.

Table 7: Minor Cover Types Found in FSA CLU Data for CROP1 Variable

Minor Cover Types	# CROP1 CLU♣s > 0	Total CROP1 Acres	# CLU♣s Usable for Training	Acres Usable for Training
Oats	64	787.0	45	427.0
Sorghum Forage	49	422.5	40	332.8
Waterbanks	5	171.1	5	171.1
Wildlife Habitat	14	116.7	12	33.7
Millet	5	52.2	3	20.4
Clover	5	30.4	2	2.4
Sunflower	2	29.4	0	0
Turn Areas	18	29.2	18	29.2
Mixed Forage	1	16.3	0	0
Home Garden	9	15.9	9	15.9
Triticale	2	5.0	2	5.0
Pulp Trees	4	1.8	4	1.8
Rye	1	0.6	1	0.6
Plums	1	0.6	0	0
Total	180	1,678.7	131	1039.9

For the 5-county test area, the only minor crop that would be added is Oats; to NASS, Sorghum Forage is just a use of Sorghum planted, while Waterbanks is covered in the NASS signature set for waste and non-agricultural areas. However, in a larger area than that of the current Project, other minor crops might be represented by enough acreage to create signatures.

Objective 3. Compare NASS segment data with FSA data for NASS segment areas.

As mentioned earlier, the CLU based ♧segments♣ were created by using ArcView to intersect the outer boundary of the 29 NASS area frame segments with the 5 county CLU shapefiles. For split CLU♣s with more than one crop type, the CLU amounts in the new segments were prorated from the original amounts based on each part♣s overall size. In a few cases, a manual review allowed determination of which part contained which crop/cover, and a proration was not necessary. Table 8 compares the two sets of segments with respect to the measurable contents. The final signature file used for classification in each case included 42 signatures of ♧extra♣ covers such as clouds, urban, water, and dense woodland; these were created manually from the full two Landsat scene area in analysis AD01.

Table 8: Comparison of FSA versus NASS Data in the 29 Segment Area

Cover	# CROP1 >0 CLU♣s / Fields		Number ♣Usable♣ CLU♣s / Fields		Statistics File # Categories		Training Pixels Available	
	FSA	NASS	FSA	NASS	FSA	NASS	FSA	NASS
Corn	43	106	31	62	11	6	5,007	4,855
Soybeans	58	143	35	75	12	10	6,131	5,991
Sorghum	32	72	12	27	1	2	949	1,699
Alfalfa	9	27	4	5	2	2	109	98
Win Wheat	8	35	4	4	6	7	681	88
Fallow	4	3	2	3	* 1	2	19	124
CRP / Idle	41	51	38	26	3	2	2,825	1,977
Grass / Past.	43	60	21	42	6	2	726	2,910
Oats	1	4	1	0	2	0	154	0
Forage S/S	1	n/a	1	n/a	* 0	n/a	22	n/a
Non Agric	n/a	163	n/a	14	n/a	6	n/a	1,167
Other Crops	n/a	6	n/a	2	n/a	1	n/a	557
Wild Hay	n/a	34	n/a	4	n/a	2	n/a	88
Woods	n/a	18	n/a	3	n/a	2	n/a	116
Farmstead	n/a	35	n/a	1	n/a	0	n/a	1
Other Hay	n/a	11	n/a	2	n/a	0	n/a	5
Total	** 240	773	149	270	44	44	16,623	20,114

* Training pixels from Fallow and Forage S/S were joined to create 1 category of ♣Other♣.
** There were 496 CLU♣s with no 578 data, a total of 736 total CLU♣s in the 29 segments.

Several categories were not present in both datasets. Some covers with different definitions were considered similar and shown in the same row, including CRP with Idle Cropland and Grass with Permanent Pasture. Removing those CLU♣s with no sign-up data, there are significantly more fields overall and more fields usable for training in the NASS segments. Looking at specific cover types, corn and soybeans are represented at about the same level in each dataset. Winter wheat and oats have more training pixels in the FSA dataset, which is encouraging for the more minor crops. The NASS dataset contains more information on Non-Agricultural areas, pasture, and hay. Only for corn does the number of categories created by clustering differ significantly; the FSA based corn is much more variable than that of the NASS segments. Interestingly enough, both sets of training pixels created the same number of classifier categories.

Even though not stated in the original objectives, one other analysis was performed to compare the two classifiers created from the 29 segments in the NASS (AD04) and FSA (AD02) datasets. Each of the 86 category signature files was used to classify the entire 73 (NASS) segment set derived from two Landsat scene areas. NASS boundaries were used to determine the classifier evaluation statistics for Stratum 11 R-squares and Kappa. Table 9 shows the results contrasted with the original AD01 ♧best♣ classifier. R-square statistics for the 29 segment classifiers are shown after outlier deletion.

Table 9 - Comparing FSA and NASS Classifiers on All 73 Segments

Cover	Regression R-squares - Stratum 11			Kappa Statistics - All Segments		
	NASS(73)	NASS(29)	FSA(29)	NASS(73)	NASS(29)	FSA(29)
Corn	0.920	0.798	0.664	86.21	58.37	48.19
Soybean	0.893	0.767	0.766	88.86	77.69	64.24
Sorghum	0.909	0.857	0.485	85.38	86.06	55.00
Alfalfa	0.804	0.741	0.255	90.27	34.63	9.45
WinWheat	0.903	0.854	0.887	97.92	76.03	48.86
All Covers	n/a	n/a	n/a	82.04	63.29	* 40.74

* Several non-crop covers are not represented by name in the FSA signature set, causing this statistic to be somewhat lower than it should be. It would probably be somewhere close to or just below the Kappa number for Corn.

Table 9 attempts to measure the signature extension capabilities of the two training sets from a 5-county area against a set from the entire 2 Landsat scene, 20 plus county area. Both are significantly reduced from the original 73-segment analysis. Again, the FSA-based data set could be significantly improved for training with a manual edit, as was done in the NASS fields.

Objective 4. Compare NASS categorized data with FSA for the entire area and not just NASS sampled areas.

Two approaches were used to analyze this objective: comparing summary data by county for major crops and counting classified pixels by CLU in a percent correct design. For the major crop by county summary approach, the variables compared were: 2001 FSA Reported (578) Acreage, the 1997 Census of Agriculture estimate, the 2001 NASS published (PEDB) county estimate, and three classified pixel based estimators from the 2001 AD01 analysis. The pixel based estimators are: the regression based county estimate, a Simple Adjusted Pixel Count Estimator (SAPCE), and a Raw Pixel Count estimator. Table 10 shows the county summary comparison for corn, soybean, and total cropland.

Table 10: County Summaries for Major Crops and Total Cropland, in Acres

County: Variable:	Gage	Jefferson	Lancaster	Saline	Seward	5-county Sum	5 county % PEDB
CORN							
FSA Rptacre1-7	71577	32694	72513	66864	70508	314156	58.28
Ag Census 97 Hv	76200	54874	90853	85699	128849	436475	80.98
SARS Regr Pltd	112114	75495	126852	106977	133346	554784	102.93
ESTPIX SAPCE	119854	72747	122427	113033	133030	561091	104.09
Raw Pixel Count	118118	71610	119410	111782	130320	551240	102.27
PEDB 2001 Pltd	115000	76000	122000	96000	130000	539000	
SOYBEAN							
FSA Rptacre1-7	86327	40713	80235	68394	63042	338711	60.11
Ag Census 97 Hv	106681	60716	102419	64059	76458	410333	72.82
SARS Regr Pltd	155126	71413	124319	101951	106249	559058	99.21
ESTPIX SAPCE	147609	77806	125153	98926	107466	556960	98.84
Raw Pixel Count	160761	84739	136304	107741	117042	606587	107.65
PEDB 2001 Pltd	133000	81500	137000	99000	113000	563500	
Total Cropland							
FSA Rptacre1-7	245386	176585	181226	228097	165271	996565	
AgCen. 97 Cropl.	342431	201576	287382	235212	251976	1318577	
*Cropland Pixels	390660	220211	328055	271420	286066	1496412	
Crops+Hay Pixels	341854	194580	285904	256996	268055	1347389	

* Includes (crops,hay,idle,fallow); excludes (woods,water,urban,pasture_other)

The FSA data for corn and soybean acreage reflects about 60 percent of the PEDB Official NASS county estimates; which is consistent with the earlier data of 48 percent overall reported via ✣578♣ signups, and expecting crops to have a higher rate than all land overall.

In order to compare the FSA data directly to the classification on a CLU by CLU basis, a percent correct approach was used. In this analysis, the FSA CLU polygons with only CROP1 > 0 are considered to be ✣ground truth♣ in the remote sensing sense, and the AD01 classification will be the variable to be evaluated. The IMG format file from the Nebraska 2001 CDL product (based on the AD01 classifier) was converted to ERSI♣s GRID format. The ArcView Spatial Analysis module was used to count the classified pixels by cover type within each CLU for the 5-county area. Table 11 shows the percent correct and commission error statistics for this comparison.

Table 11 - NASS AD01 Classifier Performance Measured Against CROP1 Only FSA CLU Polygons

Cover Type	Percent Correct	Commission Error
Corn	75.13	23.97
Soybean	77.20	21.48
Sorghum	40.22	39.30
Alfalfa	37.13	41.87
WinWheat	28.79	36.47
Other Crops	0.07	99.93
Non-crop	81.64	38.67
All Covers	71.67	n/a

Objective 5. Look at methods to improve classification accuracy with additional FSA data as input (related to item 2).

Classification accuracy can be improved in two ways: better signatures (i.e. better covering the possible spectral range of the specific cover) for the cover types you do have, and additional signatures for the cover types not represented in the original set. Table 12 contrasts the acres available for training for the entire 73 segment NASS set over 20+ counties, the 5-county FSA (with only CROP1 >0) area alone, and a projected 20-county FSA area (i.e., four times the 5-county area).

Table 12: Acres Usable For Training, Actual Versus Projected

Area: Cover Type:	NASS 20-county 73 Segments Actual	FSA 5-County Actual	FSA 20-County Projected
Corn	4,245	145,098	580,000
Soybean	4,213	153,713	616,000
Sorghum	385	34,789	140,000
Alfalfa	104	9,409	37,500
Winter Wheat	163	15,517	60,000
Oats	0	427	1,500
Other Misc. Crops	124	378	1,500
Total	9,234	359,331	1,436,500

Obviously, the FSA data set will have a much larger pool of training data available, even if much more stringent editing of this data is performed. Consider the following three sets of crop groups: major (corn and soybean), medium (sorghum and winter wheat) and minor (alfalfa, oats, and other miscellaneous crops). With respect to major crops, the usable FSA data could be sub-sampled based on field size to create more signatures than were created in the NASS 73-segment analysis. The sub-sampling of FSA data could also be tailored for areas with a sparse NASS sample due to cloud cover problems or in types of areas with very few NASS segments (stratum 20 and other non-agriculture intensive strata). The same is true of medium importance crops in this area such as sorghum and winter wheat. This would lead to more precise signatures overall for the major and medium crop types. With respect to minor crops, the alfalfa signatures could be improved and signatures created for oats (definitely) and other minor crops (possibly). In more diverse areas than seen in the pilot 5-county area, the minor crops would probably become more prevalent.

Conclusions and Recommendations

Overall, the prospect of using FSA polygons to enhance the Cropland Data Layer Program at NASS is very encouraging. There are three obstacles to overcome in using FSA CLU data as input to the CDL Program. First, a CLU polygon may have up to seven cover types, while training data for a supervised remote sensing classifier must be from one cover type. However a significant portion (50-60%) of the FSA CLU♣s have only one cover type or only a minor amount of area not in the dominant cover type; this subset of polygons can be used as training data. The second obstacle is that some CLU♣s have no ♣578♣ sign-up cover type from the farmer for the current year. For the 5-county area in 2002, only about 48% of the land area was signed up. Considering the first two points, only about one fourth of the FSA polygons would be usable for training; however, that is still a large amount of acreage. Finally, the third obstacle is the amount of work involved in conversion of CLU shapefiles to PEDITOR usable formats. The following paragraphs address the results by Objective, and offer recommendations for the future.

For Objective 1, using the NASS standard procedures and only June Area segments as input, a very good classification for the two Landsat scene area in eastern Nebraska was obtained. This was based on 73 NASS segments found in the overall 20 plus county area. A standard Cropland Data Layer (CDL) product was created and circulated to the NASS State Statistical Office in Nebraska, the Nebraska Research Initiative staff at University of Nebraska-Lincoln, and to FSA staff in headquarters and the Nebraska State office. A 5-county subset of this analysis was also done, using a reduced set of 29 segments. The reduced analysis was reasonable comparable to the original, at least for the 5-county area.

To perform Objective 2, categorizing the same 5-county area using only FSA data from the same 29 segment areas (and create more minor crop break outs if possible), another classification was created. This is a reasonable, although slightly worse, classification of the same segment areas as the NASS based 29-segment subset mentioned above, with the exception of a significant drop in the corn and alfalfa r-squares. The corn problem (and possibly alfalfa also) could probably be attributable to the lack of a manual edit on the FSA dataset. To make this more ♣fair♣ for the FSA classification, much more training data could have been gleaned from the FSA CLU dataset, but outside of the original segment boundaries. I would expect this classification to be much improved over the segment limited one based only on FSA data in segment areas. For the 5-county test area, the only minor crop that would be added is Oats; however for larger or more diverse areas, other minor crops would have signatures.

Two approaches to Objective 3 were used to compare NASS segment data with FSA data for NASS segment areas. First, the amount of training data (measured in usable pixels) and number of categories generated by clustering them were compared. Although there were 20 percent more pixels usable for training via the NASS boundaries, the number of pixels available for training was almost equal for corn and soybean. Overall, the respective clustering analyses produced approximately the same number of categories for signatures. The main exception was corn, which was much more variable in the FSA analysis, with 11 categories versus 6 for NASS. This is felt to be an editing problem, meaning that the FSA boundaries will need to go through a manual/visual edit as is already done with the NASS boundaries. Second, the two reduced set classifiers were compared to the overall ♧best♣ classifier based on all 73 segments. The FSA based training set did not stand up very well in this analysis, leading to the conclusion that it should not be used as a stand alone source for training pixels without a thorough edit process.

Objective 4, Comparing NASS categorized data with FSA for the entire area and not just NASS sampled areas, was hard to quantify. The missing data problem was evident, with corn and soybean FSA totals for the 5-county area being 58 and 60 percent of the PEDB total estimate. Approaching this from another angle, the ♧best♣ NASS classifier was measured as if the subset of FSA CLU boundaries containing only one crop (CROP1 > 0 and CROP2-7 = 0) was the ground ♧truth ♣. Percent correct for both corn and soybean was in the mid-to-high 70's range, with a commission error in the low 20 percent range. Since the CLU♣s are not arranged as segments in this analysis, no regression r-square was available. These percent correct numbers are directly comparable to those NASS would obtain using all segment data including those fields labeled as ♧bad for training♣.

For Objective 5, reviewing methods to improve classification accuracy with additional FSA data as input, several approaches are available. With respect to major and medium importance crops, the usable FSA data could be sub-sampled based on field size to create more signatures that better cover the actual crop variability. With respect to minor crops, the alfalfa signatures could be improved and signatures created for oats (definitely) and other minor crops (possibly). In more diverse areas than seen in the pilot 5-county area, the areas of minor crops available for training would probably become more prevalent.

Overall, the following actions are recommended:

⚒ Continue working in Nebraska, increasing the area covered by at least one Landsat path.

  ▬ Note: Nebraska was added to the operational NASS CDL project for 2002.

⚒ Work with FSA to increase the CLU area supplied for 2002 (up to entire state if possible).

  ▬ Revisit the investigation of minor crop potential from FSA polygon data.

  - Test the application of proposed sampling plans and edit approaches .
  - Use the FSA training data directly to improve the NE 2002 CDL product.
- Work with FSA to solve some definition problems between the NASS and FSA data.
  - Double Cropping (i.e. winter wheat then soybeans in the same field).
  - What is Forage as a cover type versus a use of the land?
    - There are variables for Sorghum Forage (SORGF), Forage Soybean/Sorghum (FORSS) and Mixed Forage (MIXFG).
  - Where does hay (non-alfalfa) fit in the FSA definition scheme?
  - Attempt to correlate non-crop FSA and NASS cover types.
    - CRP, GRASS, Turn Areas, Water banks ... vs. Permanent Pasture, Non-Agric., Waste.
- Develop a better conversion system for importing CLU data to PEDITOR formats, and add some pre-clustering analysis tools as well:
  - Label as ♧bad for training♣ any missing data / polygons.
  - Label as ♧bad for training♣ any polygons with more than one cover.
    - Allow some amount of waste, grass, etc.
  - Sample remaining good fields by crop type.
    - allow sampling by size of field or just a systematic sample.
- Develop a formal manual/visual edit process for FSA data after it is sampled for use.
- Investigate the increased variability of corn signatures in the original FSA data set.