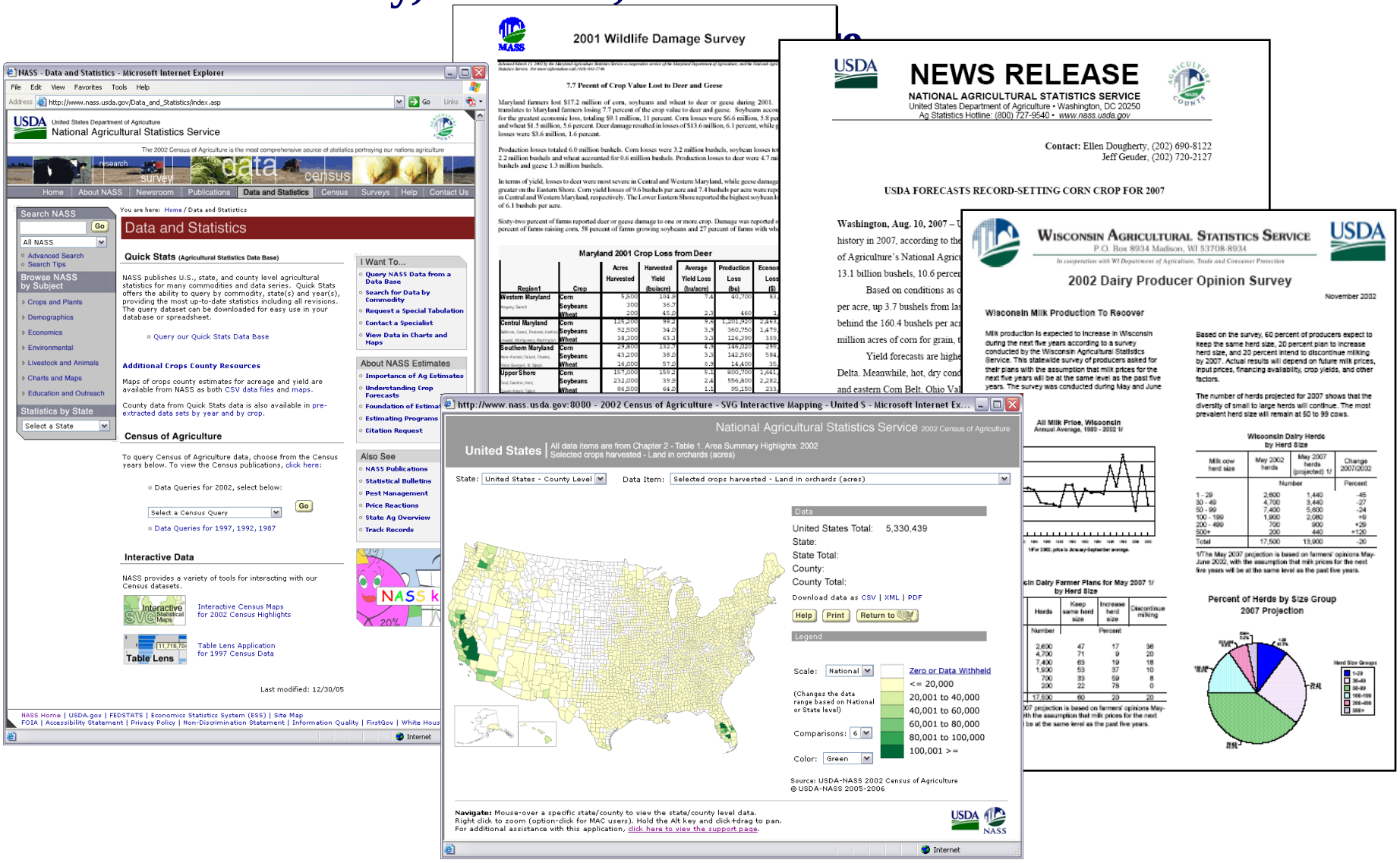
Interactive Outlier Review
and
Regression Analysis
in
Stat Studio

Bob Seffrin
Statistician

United States Department of Agriculture
National Agricultural Statistics Service
Research and Development Division
Spatial Analysis Research Section

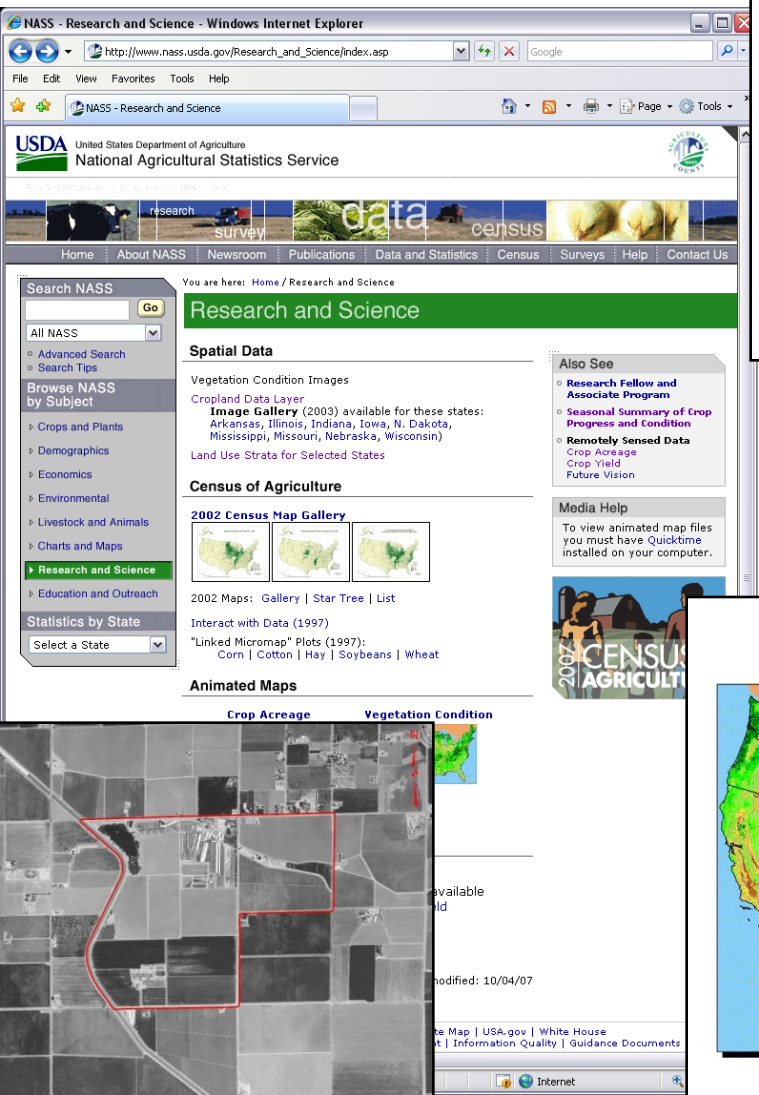
NASS Overview

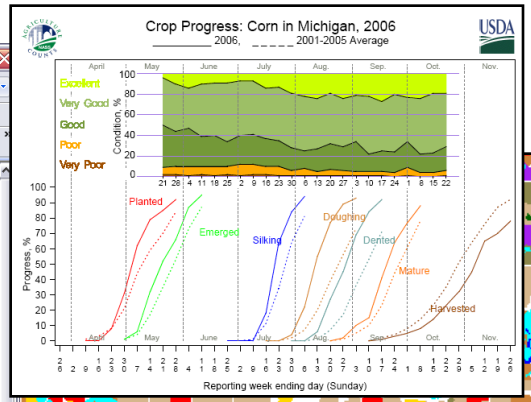
Provider of timely, accurate, and useful statistics in service to U.S.

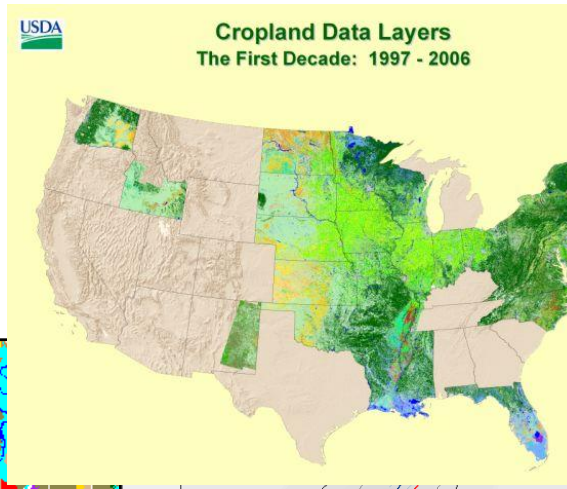


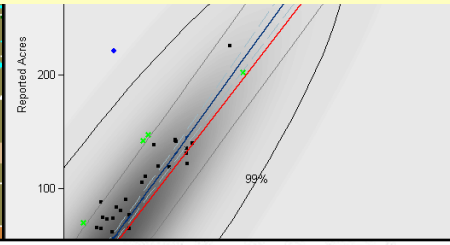
Research and Development Division

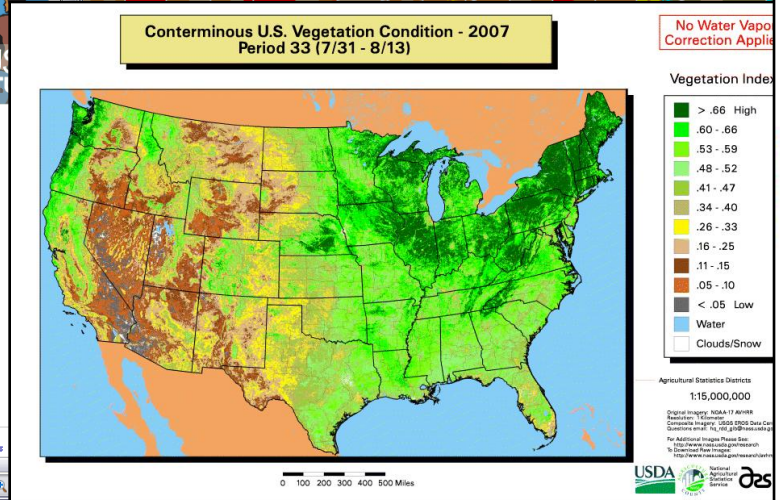
Geospatial Information Branch
Spatial Analysis Research Section

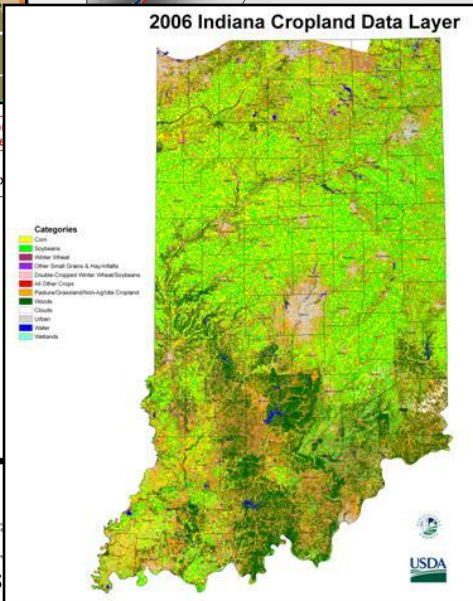




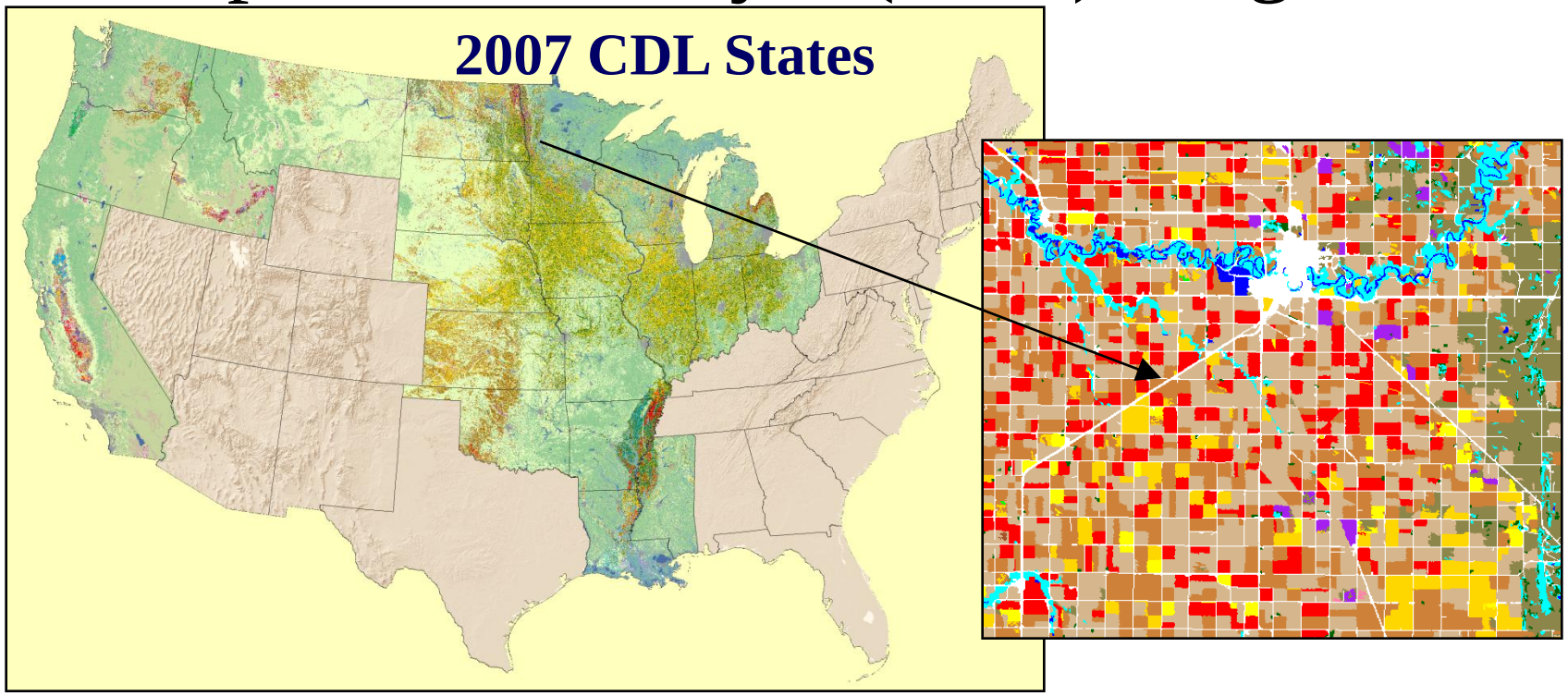








Cropland Data Layer (CDL) Program

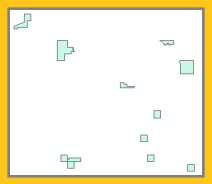


- State specific land cover classifications emphasizing row crop agriculture
  - Some regions done annually (Corn Belt, The Delta)
  - Others “one-and-done” (California, Northwest)
- Within NASS, CDL used to
  - Increase precision on survey derived acreage estimates
  - Improve county level acreage estimates

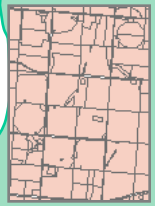
Cropland Data Layer and Acreage Estimation Processing Flow

Input Vector Data

NASS JAS segments

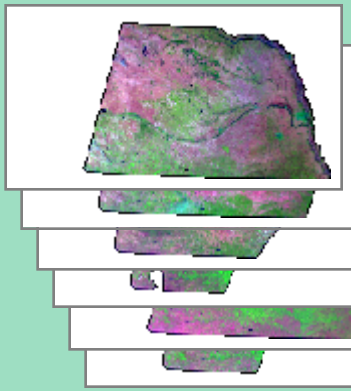


FSA CLU

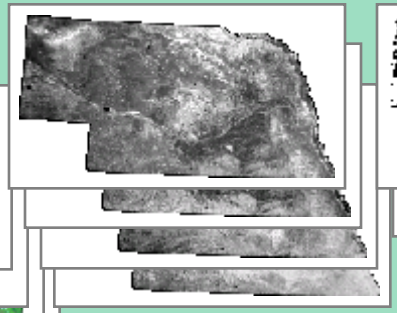


Input Raster Data

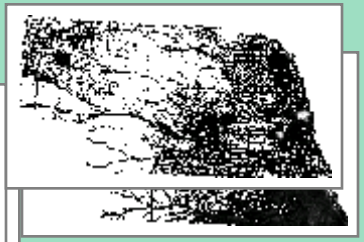
IRS Resourcesat-1 raw AWiFS summer time series



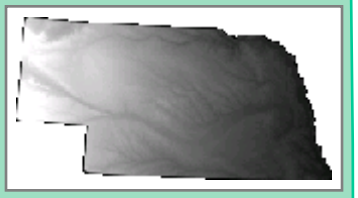
NASA Terra MODIS 16-day NDVI prior fall and summer time series



USGS NLCD circa 2001 Impervious & Canopy

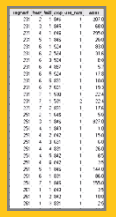


USGS NED Elevation

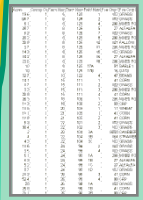


Tabular Data

JAS eData



FSA 578



All Input layers gridded to common cell size, map projection and areal extent

ESRI ArcGIS

Link and assess data sets

Ground truth

ERDAS IMAGINE®

Manages and visualizes datasets

Rulequest See5.0

Derives decision tree-based classification rules

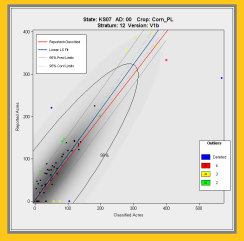
Generated rule set

ESRI ArcGIS

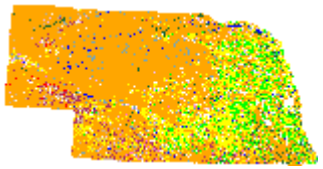
Extract JAS intersecting pixels

SAS

Customized for acreage estimation



Pixel count v. reported acreage



-- Cropland Data Layer --

datagateway.nrcs.usda.gov

State and county crop acreage statistics

Internal to NASS only

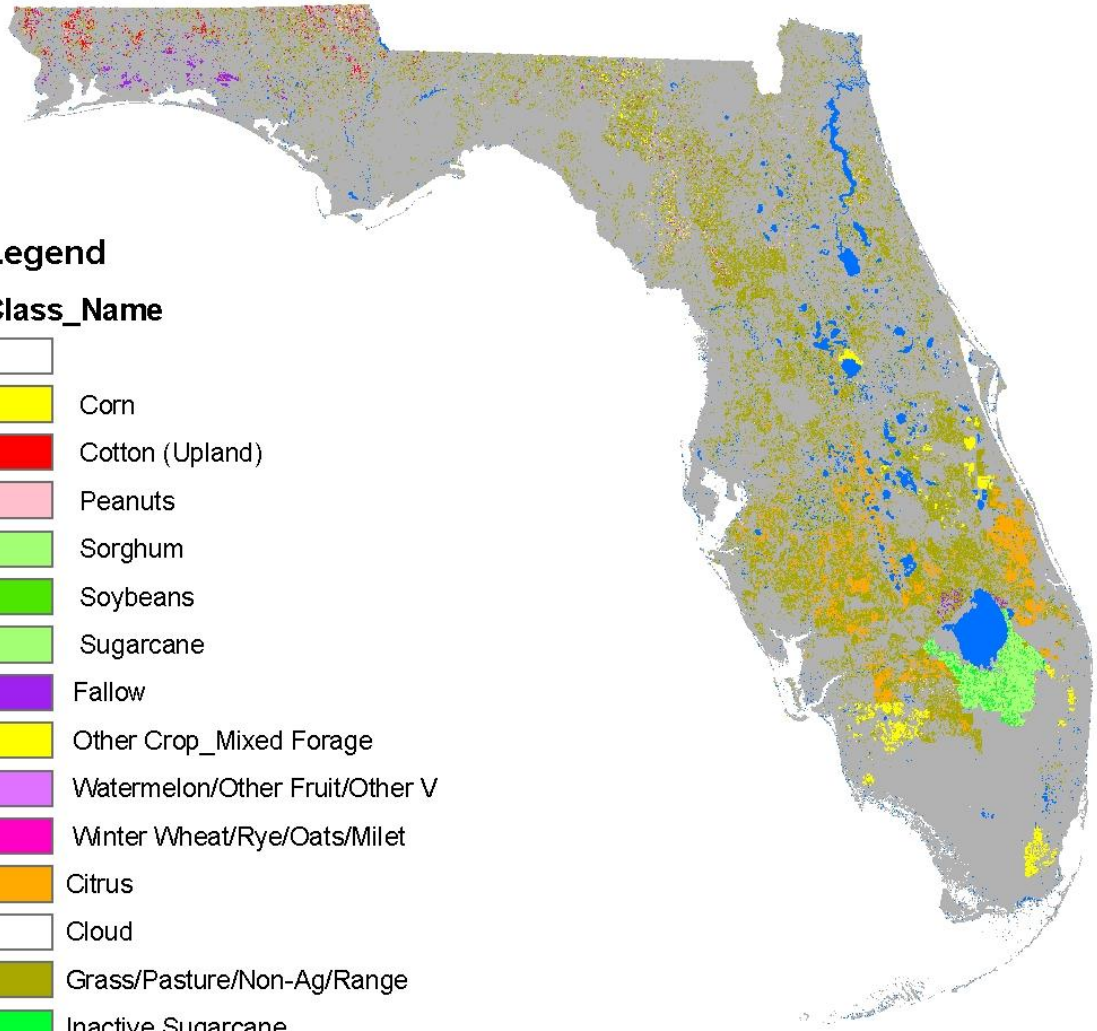
Output

Confidence Layer

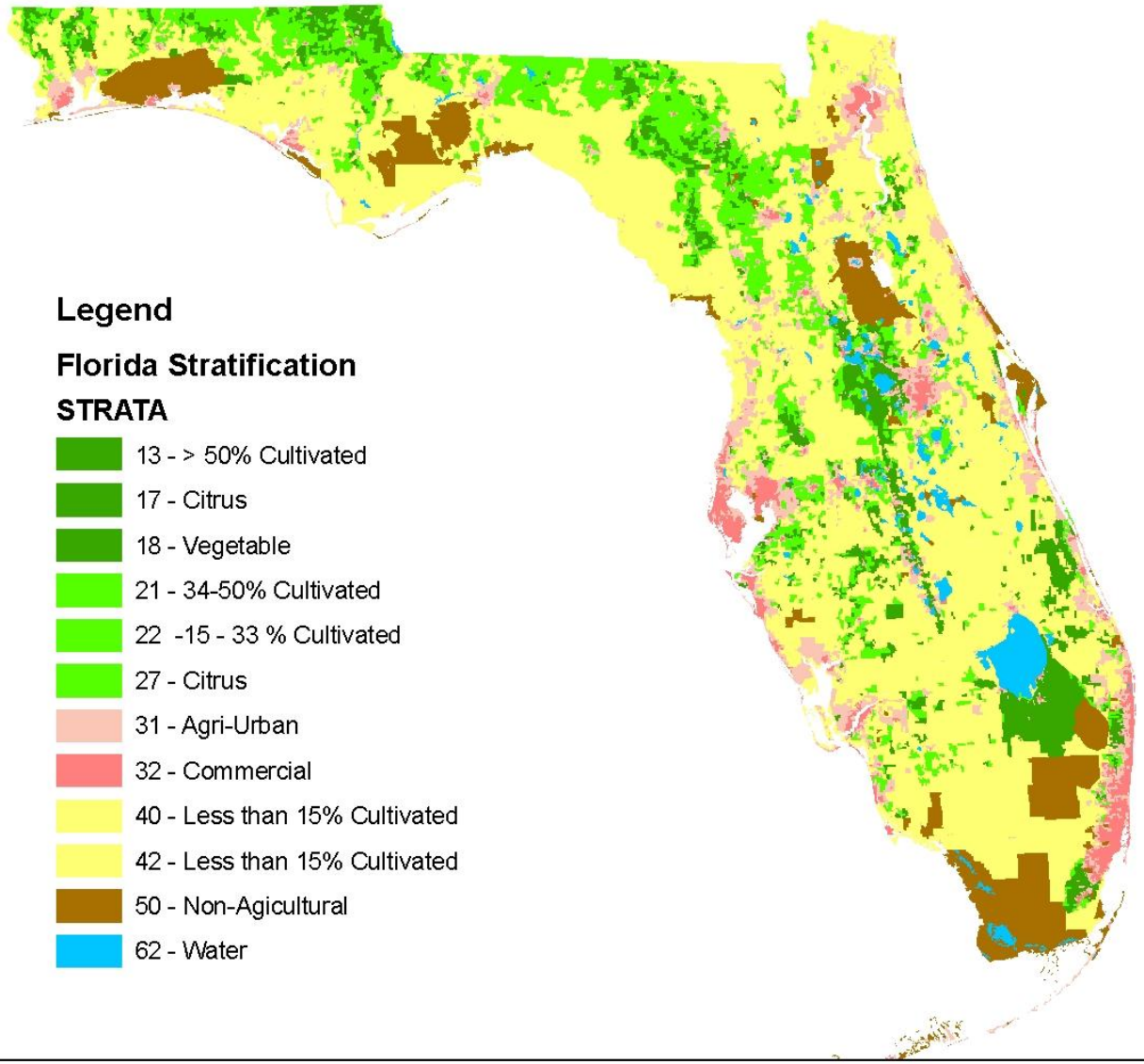
Accuracy Assessment

Diagnostics

Cropland Data Layer, 2004

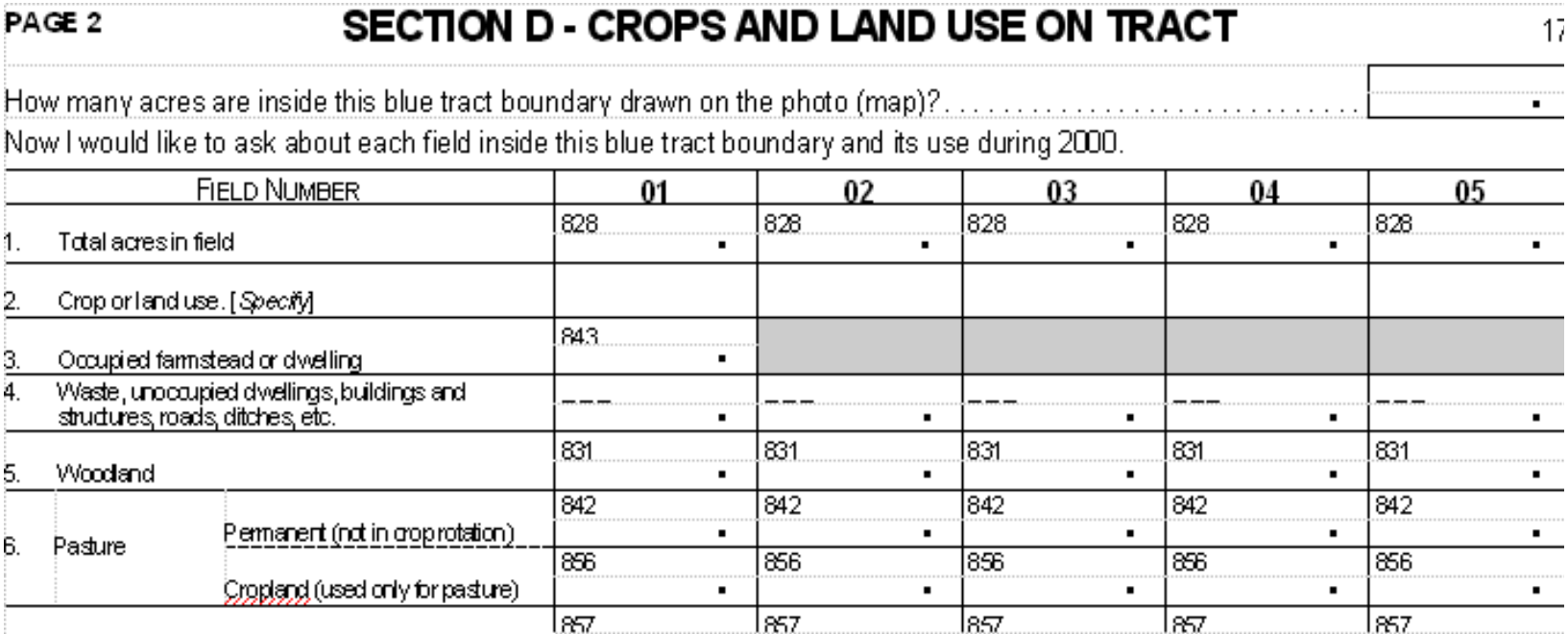


Area Frame Stratification for Florida



JAS Questionnaire

- Enumerators account for all land usage in segment
  - Draw off field location by direct observation
  - Directly link questionnaire to segment photo



PAGE 2 **SECTION D - CROPS AND LAND USE ON TRACT** 17

How many acres are inside this blue tract boundary drawn on the photo (map)? . . . . . .

Now I would like to ask about each field inside this blue tract boundary and its use during 2000.

FIELD NUMBER		01	02	03	04	05
1. Total acres in field		828	828	828	828	828
2. Crop or land use. [*Specify*]						
3. Occupied farmstead or dwelling		843				
4. Waste, unoccupied dwellings, buildings and structures, roads, ditches, etc.		___	___	___	___	___
5. Woodland		831	831	831	831	831
6. Pasture	Permanent (not in crop rotation)	842	842	842	842	842
	Cropland (used only for pasture)	856	856	856	856	856
		857	857	857	857	857

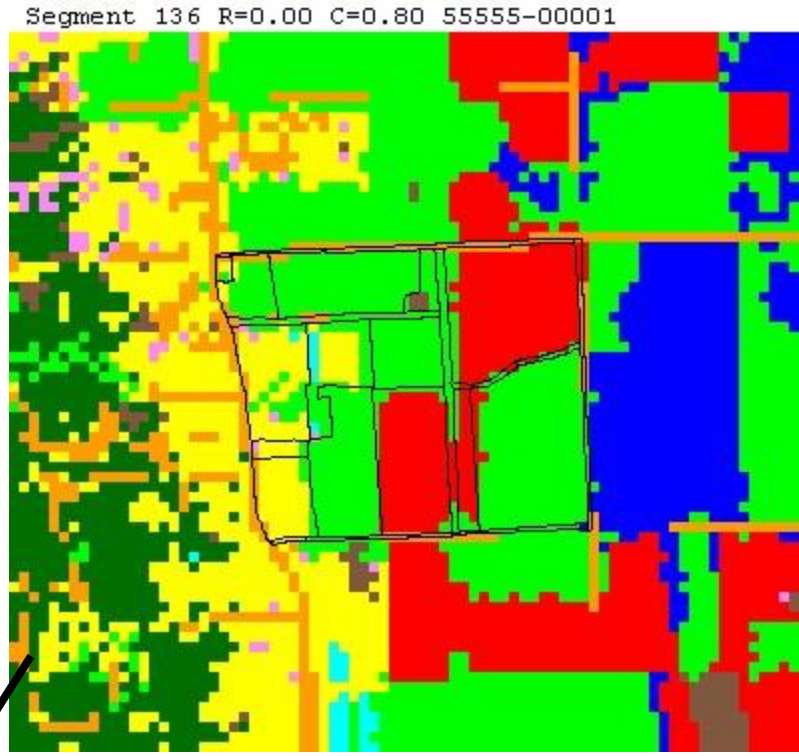
PAGE 2

SECTION D - CROPS AND LAND USE ON TRACT

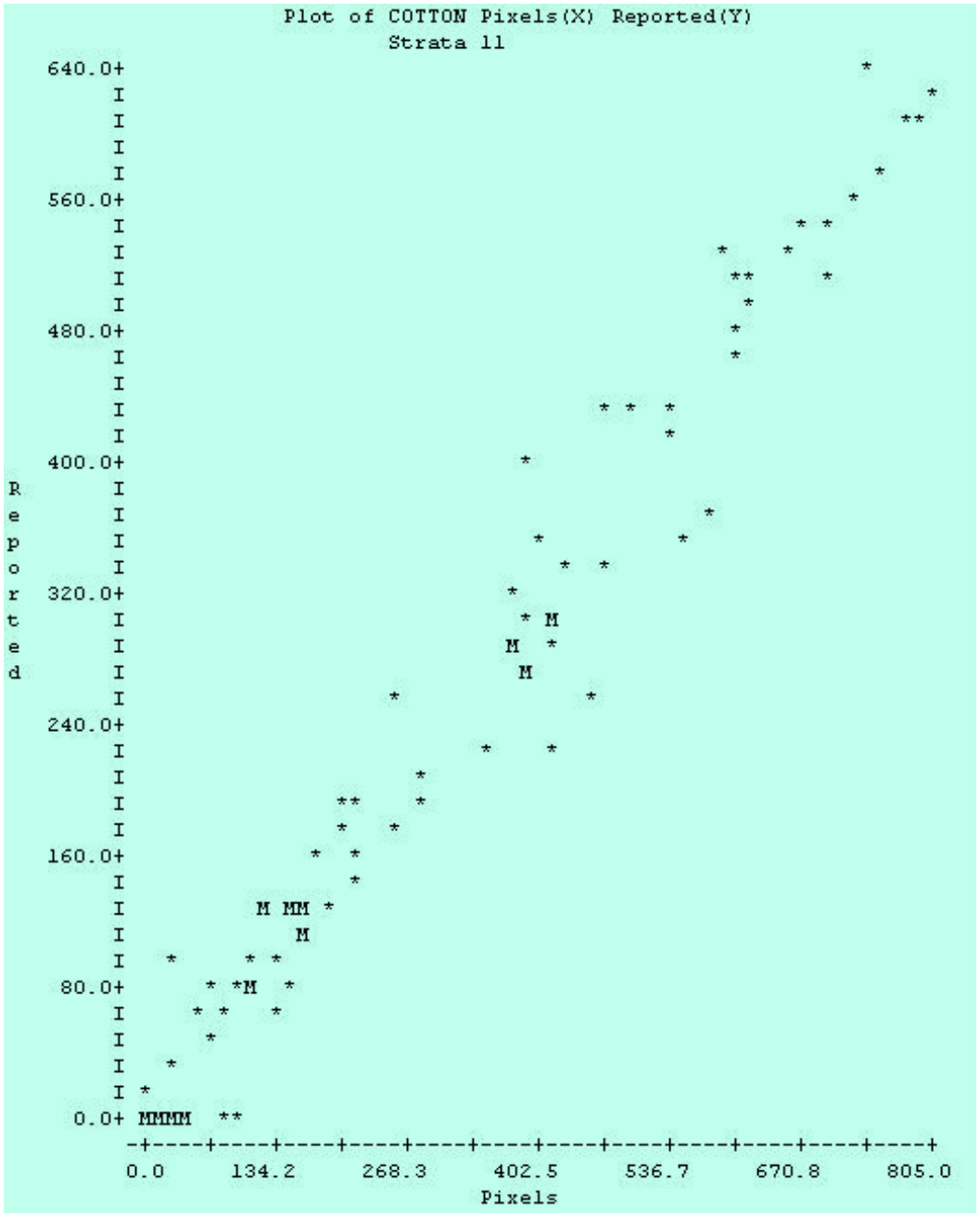
How many acres are inside this blue tract boundary drawn on the photo (map)?

Now I would like to ask about each field inside this blue tract boundary and its use during 2000.

	FIELD NUMBER	01	02	03	04	05
1.	Total acres in field	828	828	828	828	828
2.	Crop or land use. [*Specify*]					
3.	Occupied farmstead or dwelling	843				
4.	Waste, unoccupied dwellings, buildings and structures, roads, ditches, etc.					
5.	Woodland	831	831	831	831	831
6.	Pasture — Permanent (not in crop rotation)	842	842	842	842	842
	Pasture — Cropland (used only for pasture)	856	856	856	856	856
8.	Idle cropland - Idle all during 2000	857	857	857	857	857
9.	Two crops planted in this field or two uses of the same crop.	☐Yes ☐No	☐Yes ☐No	☐Yes ☐No	☐Yes ☐No	☐Yes ☐No
	[*Specify second crop or use*] Acres	844	844	844	844	844
10.	Acres left to be planted	610	610	610	610	610
11.	Acres irrigated and to be irrigated [*If double cropped, include acreage of each crop irrigated.*]	620	620	620	620	620
16.	**Winter Wheat** (*include cover crop*) Planted	540	540	540	540	540
17.	For grain or seed	541	541	541	541	541
18.	**Rye** (*include cover crop*) [*Exclude ryegrass*] Planted	547	547	547	547	547
19.	For grain or seed	548	548	548	548	548



Crop Type	Y	X
	Enumerated JAS Acres	Classified Pixels
Rice	**227.0**	**273**
Soybean	**337.0**	**541**



Pre-SAS Regression output

Table for COTTON
Strata 11

SEGMENT	PIXELS:X	REPORT:Y	ESTIMATE	RESIDUAL	HAT	RSTUD	DFFITS	COVRAT
1	391.0	265.0	303.40	-38.40	0.01	-1.76	-0.19*	0.99
2	27.0	0.0	19.92	-19.92	0.01	-0.91	-0.07	1.01
3	0.0	0.0	-1.11	1.11	0.01	0.05	0.00	1.02
4	0.0	0.0	-1.11	1.11	0.01	0.05	0.00	1.02
5	411.0	308.0	318.97	-10.97	0.01	-0.50	-0.06	1.02
6	0.0	0.0	-1.11	1.11	0.01	0.05	0.00	1.02
7	110.0	78.0	84.56	-6.56	0.00	-0.30	-0.02	1.01
8	345.0	230.0	267.57	-37.57	0.01	-1.72	-0.17	0.99
9	0.0	0.0	-1.11	1.11	0.01	0.05	0.00	1.02
10	579.0	374.9	449.81	-74.91	0.03*	-3.54*	-0.58*	0.93*
11	280.0	205.0	216.95	-11.95	0.01	-0.54	-0.05	1.01
12	97.0	0.0	74.43	-74.43	0.00	-3.47*	-0.24*	0.91*
13	0.0	0.0	-1.11	1.11	0.01	0.05	0.00	1.02
14	0.0	0.0	-1.11	1.11	0.01	0.05	0.00	1.02
15	605.0	480.0	470.06	9.94	0.03*	0.46	0.08	1.04*
16	592.0	522.0	459.94	62.06	0.03*	2.91*	0.49*	0.96*
17	0.0	0.0	-1.11	1.11	0.01	0.05	0.00	1.02
18	0.0	0.0	-1.11	1.11	0.01	0.05	0.00	1.02

Data Listing

Parameter Listing

STRATA	R-SQUARE	COEFFICIENTS				STANDARD DEVIATION	C.V.
		B(0)	STD ERR	B(1)	STD ERR	(STRATUM TOTAL)	
11	0.963	6.23	2.5860	0.8104	0.0110	22597.44	1.48%
21	0.991	1.25	1.5491	0.8184	0.0157	3735.68	3.38%
TOTAL						22904.13	1.40%

TABLE OF RELATIVE EFFICIENCIES

STRATA	DIRECT EXPANSION STANDARD DEVIATION	RELATIVE EFFICIENCY
11	116398.52	26.53
21	37671.11	101.69

CDL Estimation

Project Open

Year 2008

Month Jun

N:

State MO - Missouri

Version: v4a

Project Save

Master Categories

Grass_Resrv
Wetland_Resv
Water_Imp_St
Wildlf_Habit
Forest_Mgmtd
SkipRow_Area
Fallow,_Idle
Corn_HV
Sorghum_HV
Barley_HV
Wht_Durhm_HV
Wht_Sprng_HV

Classification File: \Acreage\MO08\Final\mo08Jun_patrick_v4a.img

file found
1a Tabulate Segment

file found
1b Tabulate Frame

file found
2a Import Seg tab.

file found
2b Import Frame tab.

file found
3 Build Regress Data

file found
Regress Seg. Summary

4 Regression

no file
Regress Param. Summary

Open SAS Log

file found
5 Estimator Selection

Close

no file
6 Build Sample

file found
7 Accumulate State

no file
Open Summary

Running: SARS Estimation on June 16, 2008
By: seffro on RDWS-RSEFFRIN

Opening project: MO - Missouri, 2008, Jun... ...ready.

IML Workshop / Stat Studio Steps

- Launch
- Generate action menu
- Run regression
- Display regression equation
- Highlight outliers
- Review outliers

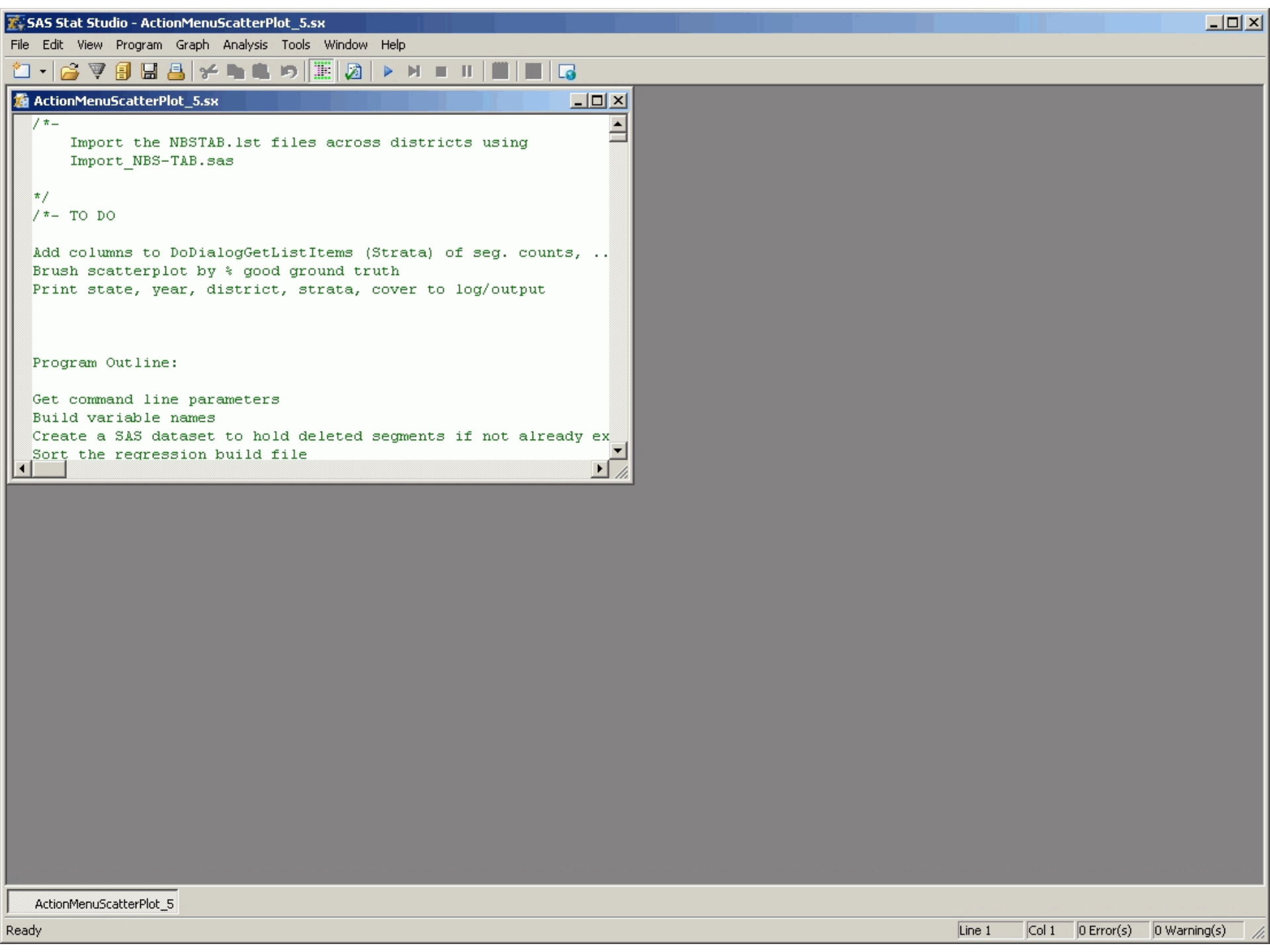
Launching IML Workshop / Stat Studio

pb_Launch_Reg:

```
 IF ComputerName = 'SARSBATCH2' THEN DO;
       Program = '"C:\...\statstudio.exe"';
       inProg    = '"C:\...\...\ActionMenuScatterPlot_5.sx"';
END;
ELSE DO;
      Program = '" C:\...\ \IMLWorkshop.exe"';
      inProg   = '"C:\...\...\ ActionMenuScatterPlot_5.iml"';
END;
rc = OPTSETN('XWAIT', 0);
rc = OPTSETN('XSYNC', 0);
 ...
```

Launching IML Workshop / Stat Studio

```
Parameters = ' -d State='   || StatePost   ||
             ' -d Year='    || Year2  ||
             ' -d Version='|| Version  ||
             ' -d LibDir='   || PathFinal ||
             ' -d inFile='   ||'Regression_Build_'|| Version ;
   ProgLaunch = Program || inProg || Parameters;
   rc = SYSTEM ( ProgLaunch );
   rc = OPTSETN('XWAIT', 1);
   rc = OPTSETN('XSYNC', 1);

RETURN;
```


SAS Stat Studio - ActionMenuScatterPlot_5.sx
File Edit View Program Graph Analysis Tools Window Help
ActionMenuScatterPlot_5.sx
/*-
Import the NBSTAB.lst files across districts using
Import_NBS-TAB.sas
*/
/*- TO DO
Add columns to DoDialogGetListItems (Strata) of seg. counts, ..
Brush scatterplot by % good ground truth
Print state, year, district, strata, cover to log/output
Program Outline:
Get command line parameters
Build variable names
Create a SAS dataset to hold deleted segments if not already ex
Sort the regression build file
ActionMenuScatterPlot_5
Ready
Line 1
Col 1
0 Error(s)
0 Warning(s)

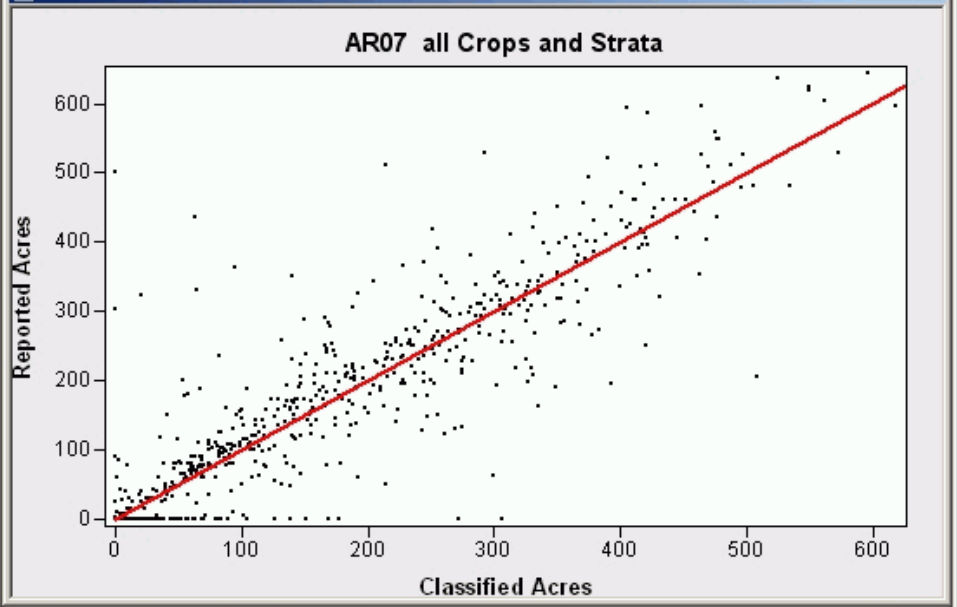
SAS Stat Studio - Regression_Build_v1b (ActionMenuScatterPlot_5)

File Edit View Program Graph Analysis Tools Window Help

ActionMenuScatterPlot_5.sx

```
/*-
    Import the NBSTAB.lst files across districts using
    Import_NBS-TAB.sas

*/
/*- TO DO

Add columns to DoDialogGetListItems (Strata) of seg. counts, ..
Brush scatterplot by % good ground truth
Print state, year, district, strata, cover to log/output



Program Outline:

Get command line parameters
Build variable names
Create a SAS dataset to hold deleted segments if not already ex
Sort the regression build file
Create data Object, set properties
```

Scatter Plot of Regression_Build_v1b (ActionMenuScatterPlot_5):2



Output1

```
Working directory:  C:\N\Estimates\Acreage\Workfiles_07_Oct\AR07
```

Regression_Build_v1b (ActionMenuScatterPlot_5):1

	9	CropName	Pixels	Stratum	AD	State
1368		Nom	Int	Nom	Nom	Nom
1	χ²	Corn_PL	0	11	00	AR
2	χ²	Corn_PL	0	11	00	AR
3	χ²	Corn_PL	0	11	00	AR
4	χ²	Corn_PL	0	11	00	AR
5	χ²	Corn_PL	0	11	00	AR
6	χ²	Corn_PL	0	11	00	AR
7	χ²	Corn_PL	0	11	00	AR
8	χ²	Corn_PL	0	11	00	AR
9	χ²	Corn_PL	0	11	00	AR
10	χ²	Corn_PL	0	11	00	AR
11	χ²	Corn_PL	0	11	00	AR
12	χ²	Corn_PL	0	11	00	AR
13	χ²	Corn_PL	0	11	00	AR
14	χ²	Corn_PL	0	11	00	AR
15	χ²	Corn_PL	0	11	00	AR

ActionMenuScatterPlot_5

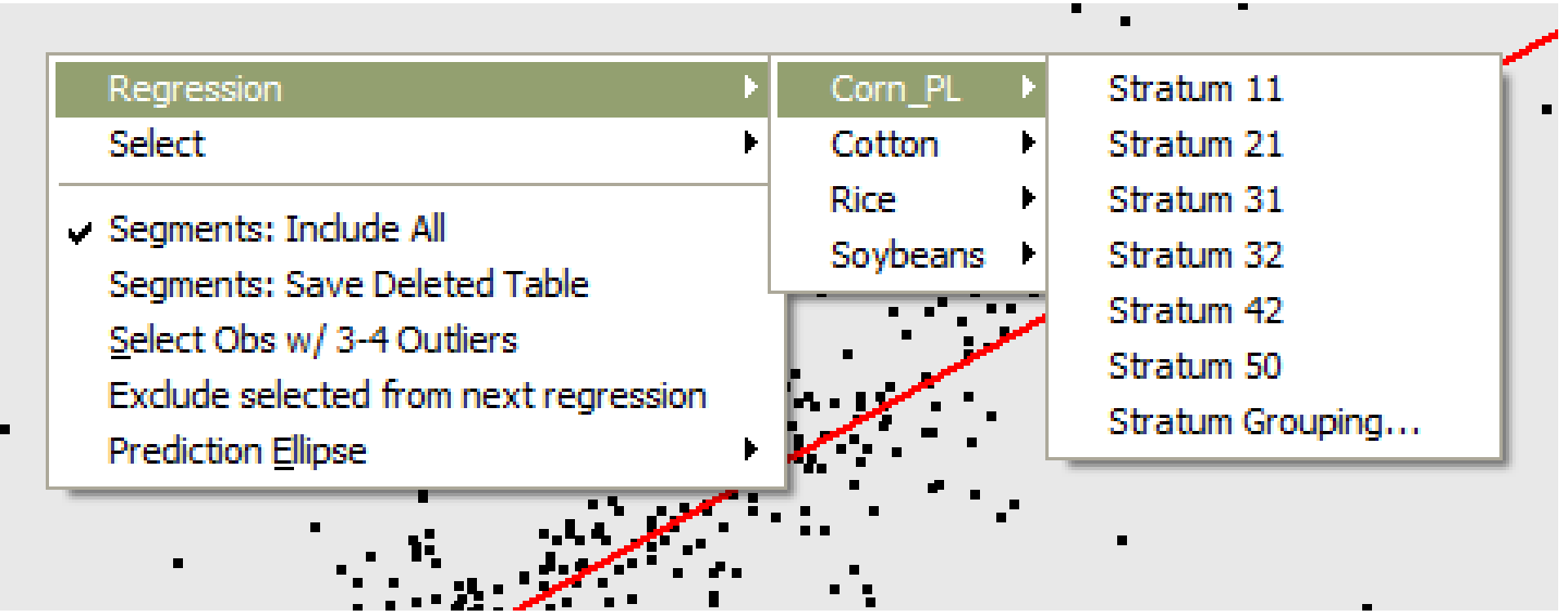
Ready 0 Error(s) 0 Warning(s)

IML Workshop / Stat Studio Steps

- Launch
- Generate action menu
- Run regression
- Display regression equation
- Highlight outliers
- Review outliers

Activating Action Menu

F11



Generating the Action Menu

Create text like:
"Regression\nState AR\nDistrict 01\nCorn\nStrata 11"J

```
uCrop     = UNIQUE( allCrop );      { also strata, state, district }
cntCrop   = NCOL( uCrop ) ;         { also strata, state, district }

TotMenuItems = cntState*cntDist*cntCrop*cntStrat;
Counts    = cntState || cntDist || cntCrop || cntStrat;
Permies   = J( TotMenuItems, NCOL(Counts), 0 );
Row       = 0;
```

Generating the Action Menu

Build indexes of possible combinations

```
DO i1 = 1 TO  Counts[1];
  DO i2 = 1 TO  Counts[2];
    DO i3 = 1 TO  Counts[3];
      DO i4 = 1 TO  Counts[4];
        Row =Row + 1;
        IF Counts[1] = 1 THEN Permies[Row,1]=0; ELSE Permies[Row,1]=i1;
        IF Counts[2] = 1 THEN Permies[Row,2]=0; ELSE Permies[Row,2]=i2;
        IF Counts[3] = 1 THEN Permies[Row,3]=0; ELSE Permies[Row,3]=i3;
        IF Counts[4] = 1 THEN Permies[Row,4]=0; ELSE Permies[Row,4]=i4;
      END;
    END;
  END;
END;
```

Generating the Action Menu

Define the root of menu, create arrays to hold results

```
sMenuBase_Reg  = "Regression\n"J;
sCode_Reg      = 'RUN OnLSRegression; ';
sCode_Group    = 'Group...';
LengthFiller = "                                                    ";
aMenuStr    = J( TotMenuItems, 1, sMenuBase_Reg + LengthFiller );
```

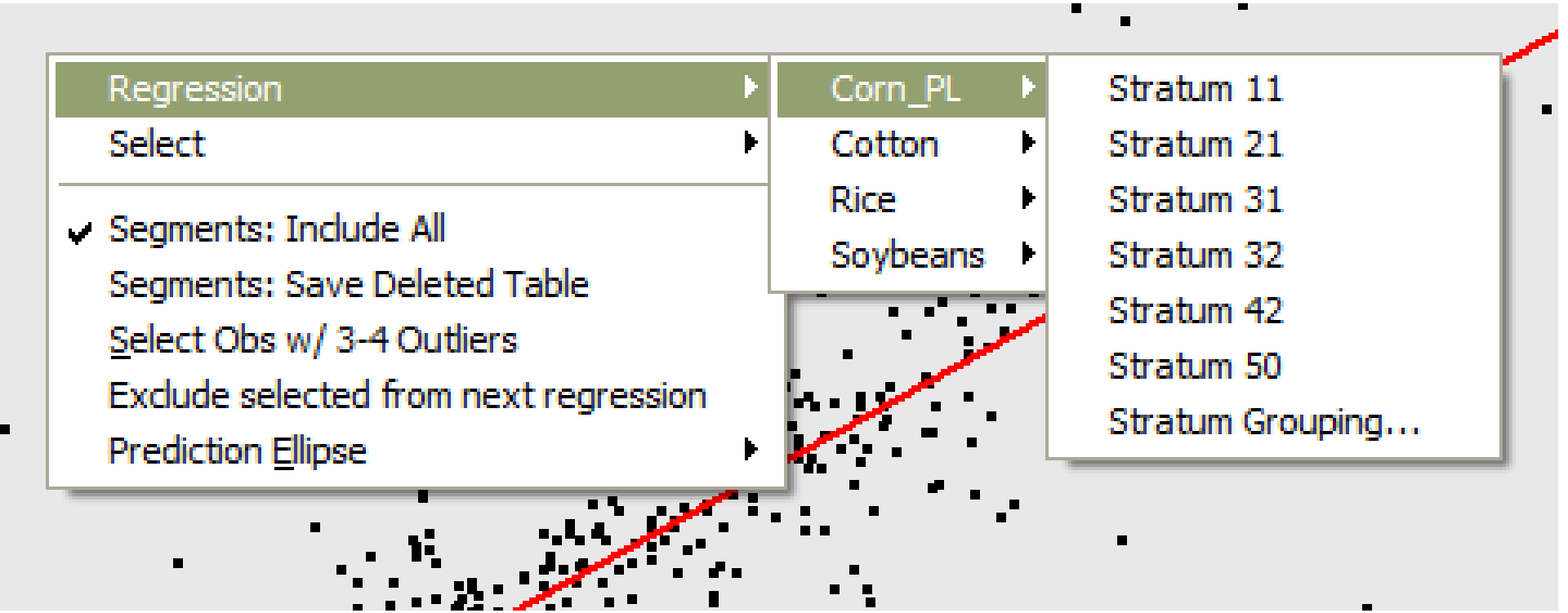
Generating the Action Menu

Apply the root of menu, cascading menues to Plot

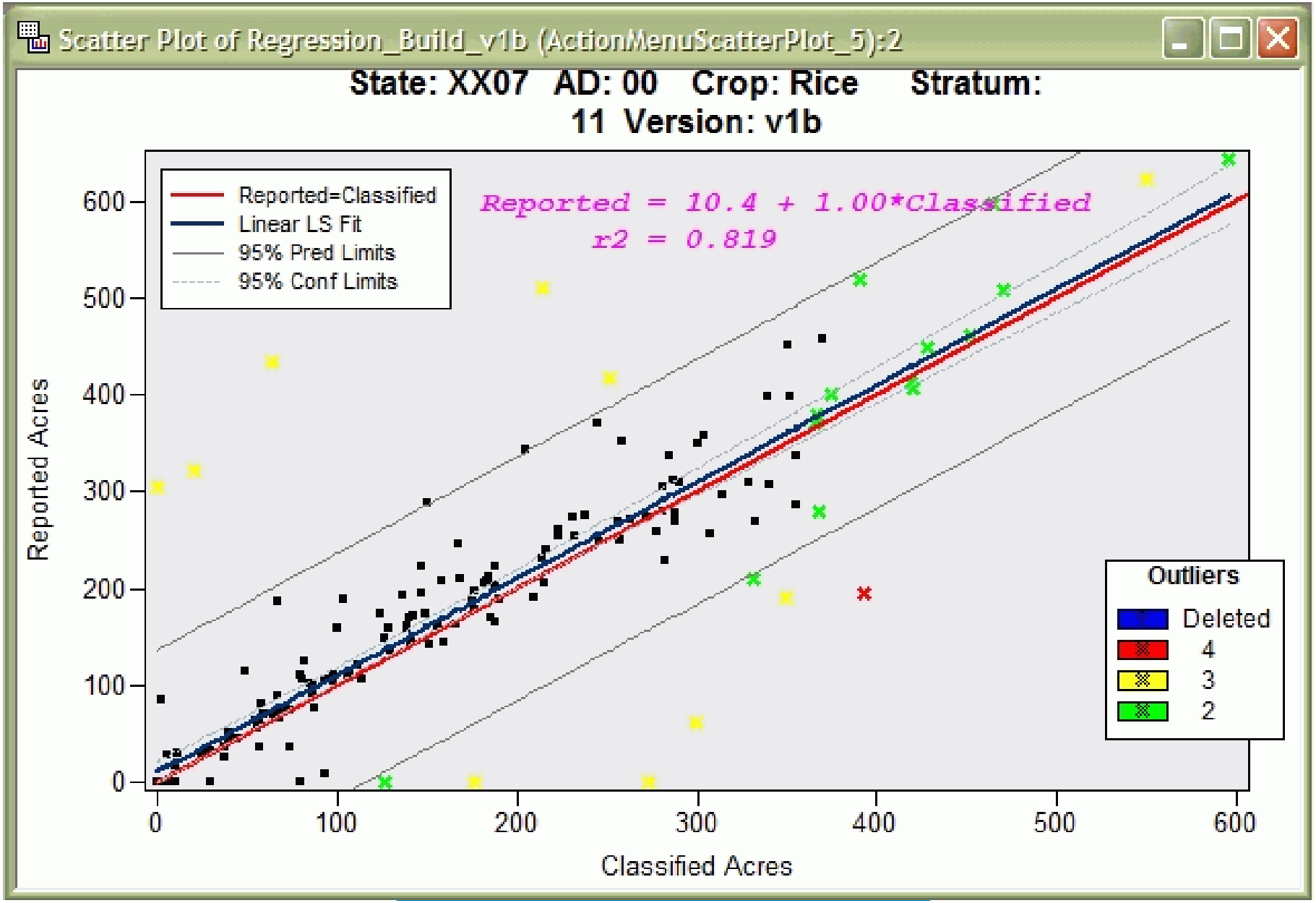
```
plot.AppendActionMenuItem( aMenuStr[1], aMenuCodeSel[1] +
                                        aMenuCodeReg[1] );

 /*- Append rest of menus to roots */
 DO i = 2 TO TotMenuItems;
    plot.AppendActionMenuItemToGroup( aMenuStr[1],aMenuStr[i]
                                     ,aMenuCodeSel[i]+Code_Reg );
 END;
```

Generating the Action Menu

Create text like:
"Regression\nCorn_PL\nStrata 11"J



Using the Action Menu



IML Workshop / Stat Studio Steps

> Launch
> Generate action menu
> Run regression
> Display regression equation
> Highlight outliers
> Review outliers

Run the Regression

```
SUBMIT< ...some variables to pass, SAS language will
        see as macro variables...> ;
  ODS OUTPUT
    ANOVA                = oAnovaFits
    Fitstatistics        = oFitstatistics
    ParameterEstimates = oParameterEstimates ;
```

Run the Regression

```
proc reg  data=RegIn   TABLEOUT ;
 &ModelStat : model &yVarName = &RegXVarNames / cl
       ADJRSQ AIC BIC CP EDF GMSEP JP MSE …;
       &WhereClause
       output out = RegOut          p = &predName
       residual   = &residName  lclm  = &lclmName
       uclm       = &uclmName     lcl = &lcliName
       ucl        = &ucliName      H  = &Hat
       rStudent   = &rStudent  DFFITS= &DFFITS
       CovRatio   = &CovRatio;
  quit;
```

Run the Regression

```
PROC SQL NOPRINT;
    CREATE TABLE FileStat <… query to merge
            regression parameters from ODS tables…>

ENDSUBMIT;
```

IML Workshop / Stat Studio Steps

> Launch
> Generate action menu
> Run regression
> Display regression equation
> Highlight outliers
> Review outliers

Display Regression Equation

```
Reported = 10.4 + 1.00*Classified
      R2 = 0.819
```

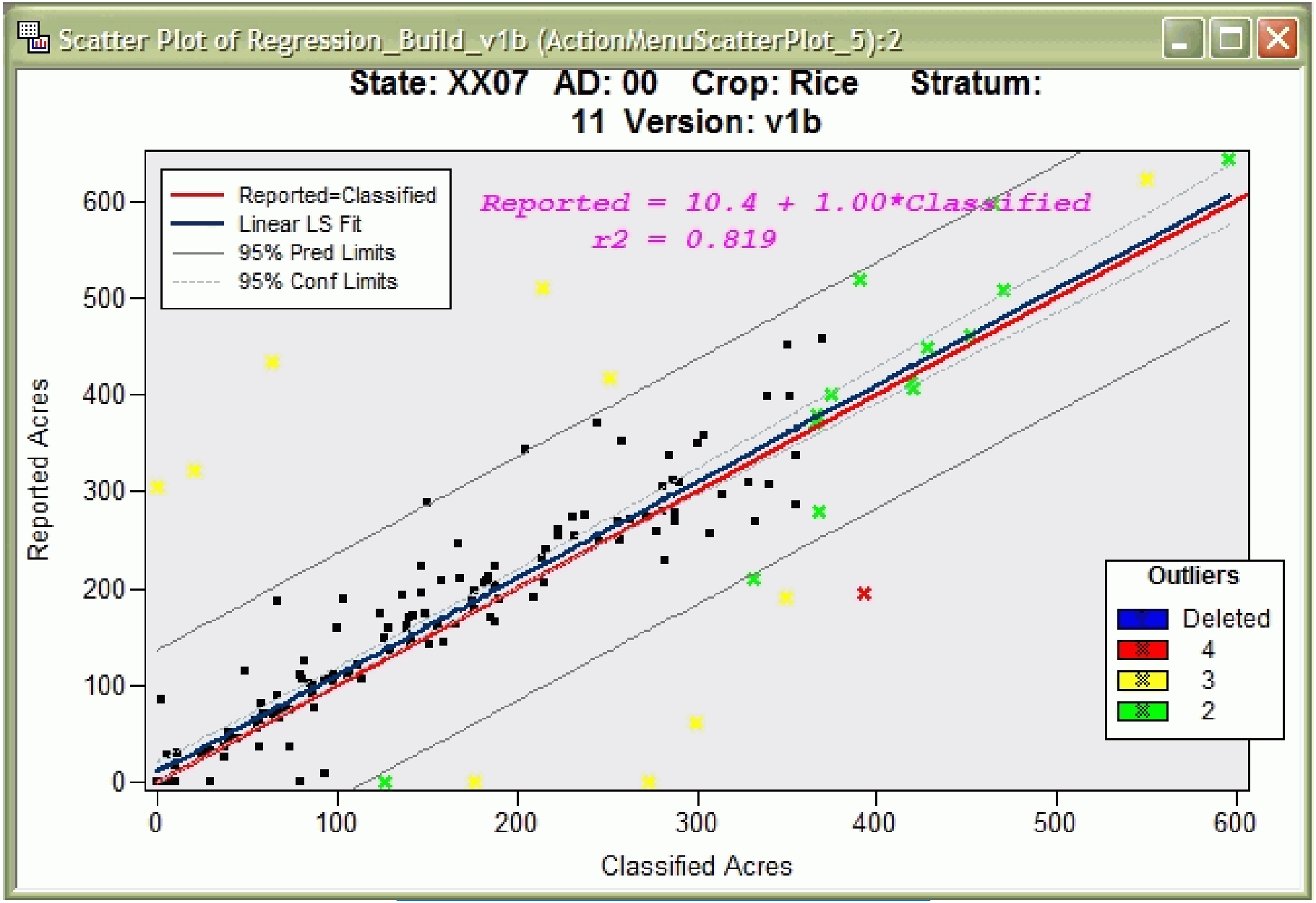
Display Regression Equation

```
plot.DrawRemoveCommands("Regress Equation");

declare DataObject RegParms;
RegParms =DataObject.CreateFromServerDataSet(FileStat)
RegParms.GetVarData( "Intercept", Reg_Intercept);
RegParms.GetVarData( "Slope",     Reg_Slope);
RegParms.GetVarData( "R_Square",  Reg_r2);
Reg_Eq = CONCAT( 'Reported = ‘
        ,STRIP(PUTN( Reg_Intercept, '4.2' ))
        , ' + ‘,STRIP(PUTN( Reg_Slope, '4.2' ))
        , "*Classified \n\r2 = "J
        ,STRIP(PUTN( Reg_r2,  '5.3' )) ) ;
```

Display Regression Equation

```
plot.DrawBeginBlock( "Regress Equation" );
    plot.DrawPushState();
    plot.DrawResetState();
        plot.DrawSetTextTypeface( "Courier New" );
        plot.DrawSetTextStyle( STYLE_BOLDITALIC );
        plot.DrawSetTextColor( MAGENTA );
        plot.DrawSetTextSize( 11 );
        plot.DrawSetTextAlignment( ALIGN_LEFT, -1 );
        plot.DrawText( 30, 93, Reg_Eq );
    plot.DrawPopState();
plot.DrawEndBlock();
```

Display Regression Equation



IML Workshop / Stat Studio Steps

> Launch
> Generate action menu
> Run regression
> Display regression equation
> Highlight outliers
> Review outliers

Highlight Outliers

Define thresholds

```
PCR  = 2;
dobjOut.GetObsNumbersInAnalysis( AnalyCnt );
FN          = NROW( AnalyCnt );
HatCR       = 2*PCR/FN;
CovCR       = 6/FN;
DffCR       = 2*SQRT( PCR/FN );
RstCR       = 2.0;
```

Highlight Outliers

Tally outliers

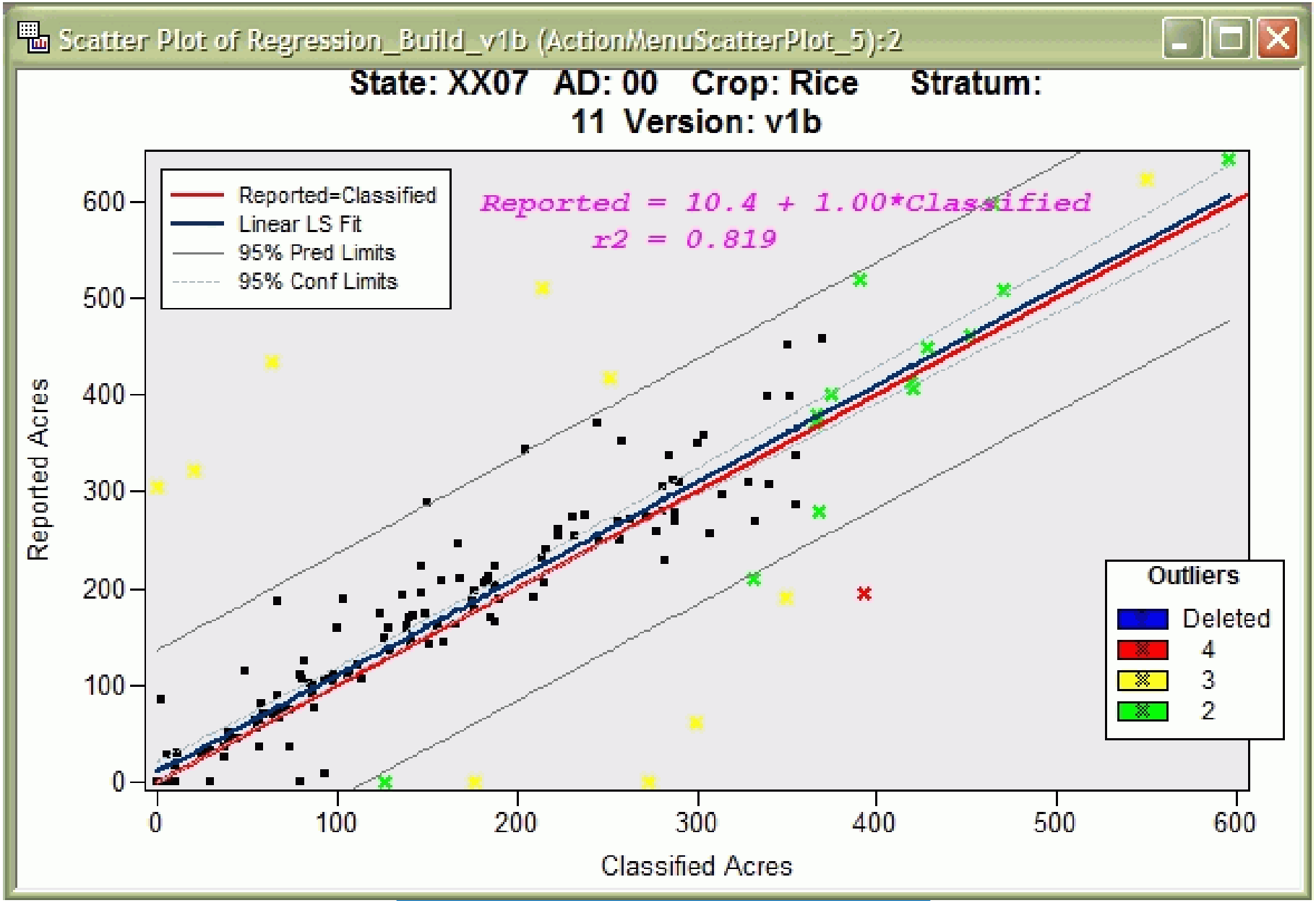
```
dobjOut.GetVarData( "Hat"    , Hat );
IF Hat[j] > HatCR THEN DO;
        OutLier[j] = OutLier[j]+1;
        OutIndi[j] = OutIndi[j]+1;
END;  {for all outlier indicators}
dobj.AddVar("OutCount", "Outlier Count,1-4“, OutLierTemp);

Out4          = LOC(OutLierTemp=4);
OutSel34      = LOC(OutLierTemp>2);
```

Highlight Outliers

Change marker and color

```
IF NCOL(Out234)>0 THEN
        plot.SetMarkerShape( Out234, MARKER_X );

IF NCOL(Out4) > 0 THEN DO;
        plot.SetMarkerColor( Out4, RED );
END;
```

Highlight Outliers



Highlight Outliers

Table: outliers, counts, colors

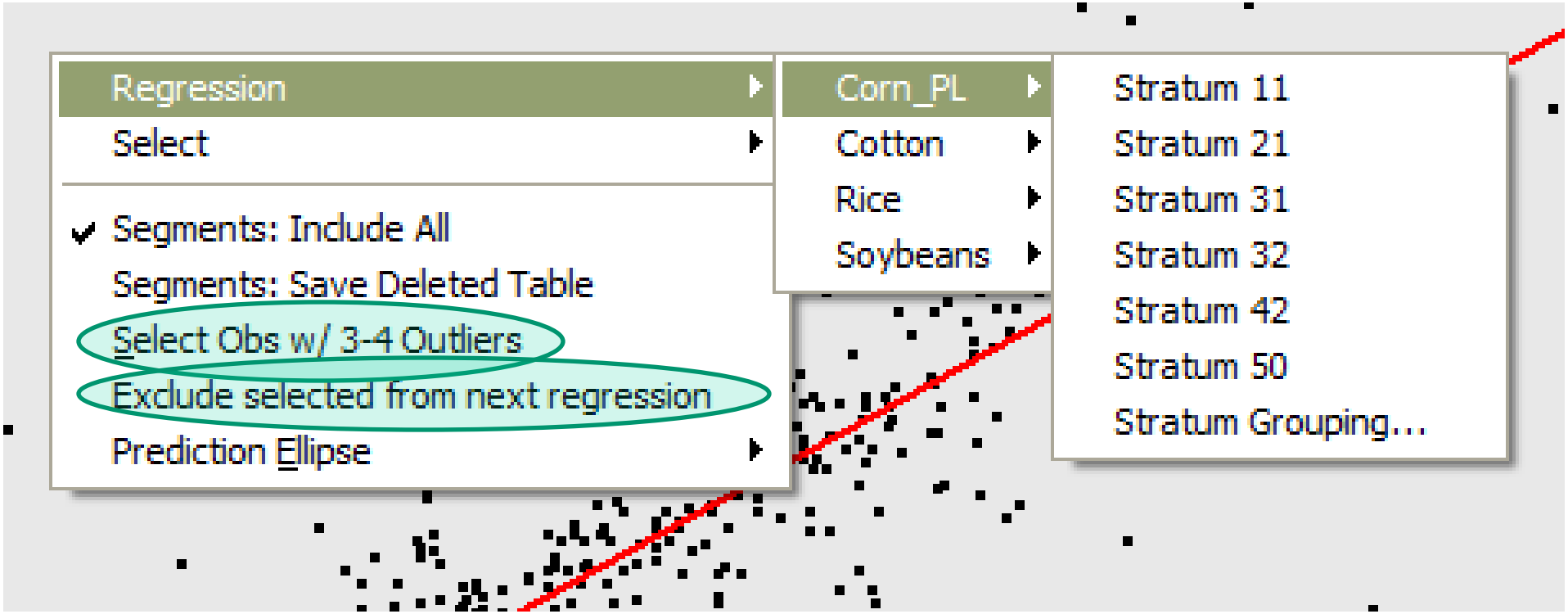
Regression_Build_v1b (ActionMenuScatterPlot_5):1

		CropName	Stratum	Segment	Hat	rStud	DFFITS	CovRat	OutCount	OutIndicator
		Nom	Nom	Nom	Int	Int	Int	Int	Int	Int
141	PC	Rice	11	141	0.01726	0.57964	0.07682	1.02347	1	1000
26	PC	Rice	11	26	0.01761	-0.437	-0.0585	1.02512	1	1000
157	PC	Rice	11	157	0.0177	-1.2693	-0.1704	1.01263	1	1000
160	PC	Rice	11	160	0.01885	-0.0985	-0.0136	1.02805	2	1001
71	PC	Rice	11	71	0.01895	0.04721	0.00656	1.02821	2	1001
3	PC	Rice	11	3	0.01913	-1.5647	-0.2185	1.00678	2	1010
140	PC	Rice	11	140	0.01941	1.20591	0.16964	1.01577	1	1000
110	PC	Rice	11	110	0.01987	0.24077	0.03428	1.02868	2	1001
79	PC	Rice	11	79	0.02191	1.89594	0.28377	0.99972	2	1010
4	PC	Rice	11	4	0.02221	-3.3764	-0.5089	0.93614	4	1111
81	PC	Rice	11	81	0.02581	-0.249	-0.0405	1.03492	2	1001
2	PC	Rice	11	2	0.02592	-0.3723	-0.0607	1.03434	2	1001

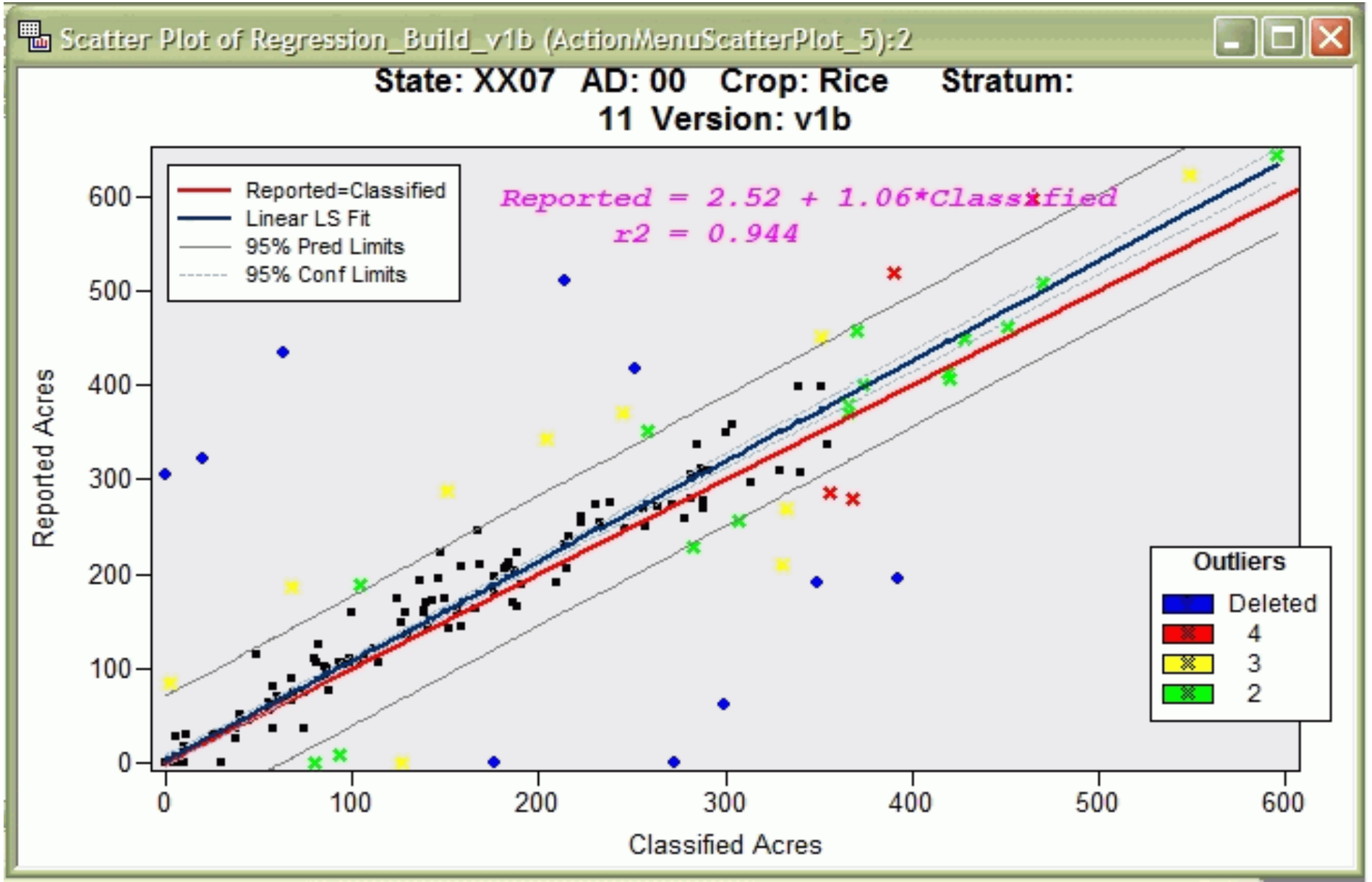
IML Workshop / Stat Studio Steps

> Launch
> Generate action menu
> Run regression
> Display regression equation
> Highlight outliers
> Review outliers

Review Outliers



Review Outliers



Stat Studio Highlights

-Interactive
-Linkage between objects
-Access to rest of SAS through SUBMIT
-Can manage graphics as named Blocks
-Power of IML
-Flexible menu

Interactive Outlier Review
and
Regression Analysis
in
Stat Studio

Bob Seffrin
Statistician

United States Department of Agriculture
National Agricultural Statistics Service
Research and Development Division
Spatial Analysis Research Section