Paper 358-2013

An Innovative Approach to Integrating SAS® Macros with GIS Software Products to Produce County-Level Accuracy Assessments

Audra Zakzeski, National Agricultural Statistics Service, U.S. Department of Agriculture;

Robert Seffrin, National Agricultural Statistics Service, U.S. Department of Agriculture

ABSTRACT

The National Agricultural Statistics Service produces an annual geospatial informational dataset called the Cropland Data Layer over the United States detailing the land cover over each state while focusing on the vast array of crops grown during the months of April through October. While calculating an accuracy assessment of the land cover over an entire state is a relatively simple process, calculating an accuracy assessment down to a county or crop specific level can be extremely time consuming and arduous. To simplify the process an innovative SAS 9.2 program was created integrating the efficiency of the SAS Macro language with the geospatial analytical capabilities of the GIS program ERDAS Imagine. The procedure is operated using a SAS/AF interface allowing analysts to easily investigate county level information.

INTRODUCTION

The National Agricultural Statistics Service (NASS) of the United States Department of Agriculture (USDA) has been producing a geo-referenced, crop-specific, land cover map called the Cropland Data Layer (CDL) for many years. For many of the more agriculturally intensive states in the country there are over a decade worth of CDL products available. A 48 contiguous state national product is available annually for 2008 to the present. The CDL is comprised of farmer reported survey data and satellite images of the land taken to measure plant density and surface reflectance in order to create unique spectral profiles for each type of land cover. A decision tree of the land cover profiles is used to classify every unknown pixel in a state.

All years of the CDL are free and available to the public therefore the possible uses are vast and stretch among many disciplines. Internally, analysts are responsible for producing multiple CDLs throughout a growing season, typically April through October, for multiple states depending on the types of agriculture grown. Each CDL is used to create crop acreage estimates defined by running a linear regression on a value of crop acreage measured by an independent USDA survey, the June Agricultural Survey, and the corresponding number of pixels for each crop classified in the CDL. Figure 1 depicts an area of Mead County, Kansas. The left image is a satellite image taken on June 23, 2012 displayed in false color infrared where plants with higher chlorophyll content, greener to the naked eye, are displayed as bright red in the image. The figure on the right is of the same area in Kansas of the CDL. Each 30 meter pixel within the CDL represents a different land cover type: yellow for corn, green for soybeans, dark brown for winter wheat, light brown for fallow land, orange for sorghum, and pink for alfalfa. When you compare the images side by side you can see the fields displayed as bright red in the satellite image generally correspond with classified corn fields, yellow pixels, in the CDL.

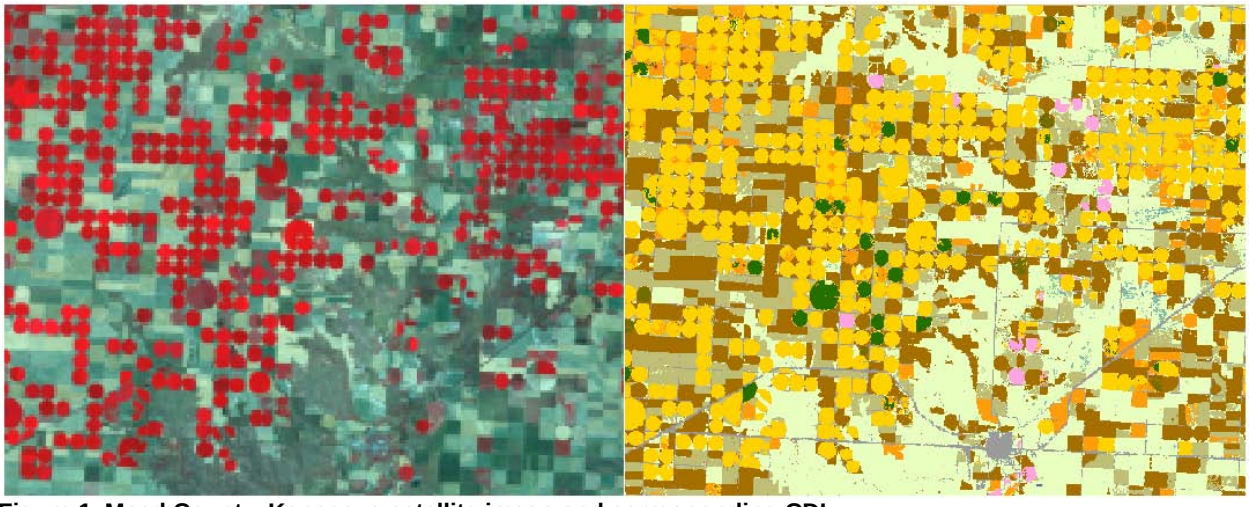

Figure 1. Mead County, Kansas; a satellite image and corresponding CDL

Since there is a wide range of land cover types, over 100, evaluating the accuracy of a CDL can be a challenge. For each type of land cover of interest (corn, soybeans, cotton, water, forest, etc) there is a measure of omission, the percentage of CDL pixels of certain land cover type that were mistakenly classified to other categories and a measure of commission, the percentage of CDL pixels of other land cover types that were mistakenly categorized to the land cover type of interest. The challenge of evaluating the CDL accuracy can become even more complex when each CDL is broken down to the individual counties in a state. Being able to see whether a crop in one county was classified more accurately than in another county can aid the analyst in further processing attempts. In order to effectively display such a vast array of accuracy measures and still maintain a geographical reference a series of batch files and SAS Macros were built to create in depth dashboards that give analysts a one page glance at state and county accuracy assessments for each crop.

CREATING A BATCH FILE TO RUN ERDAS IMAGINE MODEL

NASS uses ERDAS Imagine to analyze satellite images and create each CDL. After a CDL is created it is necessary to run accuracy assessments within ERDAS Imagine comparing the finished CDL to an independent data source of farmer reported data. In order to automate the accuracy assessment calculations a series of %LET statements, %PUT statements, and predefined macro variables can be used to create a batch file (.bat) calling ERDAS Imagine. The statements are used to define specific file paths for each of the model components and write the text for a batch file.

```
%LET DirEst = File path for Estimate file;
%LET DirCnty = File path for state/county outlines;
%LET IMGWork = ERDAS Model .exe file path;
%LET Model = Preexisting ERDAS .pmdl model file path;
%LET Zone1 = &DirCnty/Specific State Outline;
%LET Zone2 = File path for CDL File #1;
%LET Class = File path for CDL File #2;
%LET MTXout = File path for preexisting matrix file (.mtx);

%LET IMGlaunch = &IMGwork &Model –s –m &Zone1 &Zone2 &Class &MTXout;

OPTIONS NOXWAIT XSYNC;
DATA _NULL_;
FILE "File location to be created batch file" LRECL=600;
    PUT 'Set Imagine_Batch_Run=1";
    PUT 'ECHO off ';
    PUT 'TITLE ' "&StYr2.  &Tabulated";
    PUT 'COLOR 79 ';
    PUT 'ECHO Mosaic running from N:\Estimates\Acreage\Temp\Tabulate.bat";
    PUT 'TIME /t';
    PUT %unquote(%str(%')&IMGLanch %str(%));
RUN;
```

Running the code listed above writes a batch file that when executed will run the preexisting tabulation model (&Model) with ERDAS Imagine. The resulting batch filed produced by this code is displayed in Output 1.

```
SET IMAGINE_BATCH_RUN=1
ECHO off
TITLE NC12 CDL
COLOR 79
ECHO Mosaic running from N:\Estimates\Acreage\Temp\Tabulate.bat
TIME /t
"C:\ERDAS\ERDAS Desktop 2011\bin\Win32Release\modeler.exe"
"N:/Estimates/Calibration/Progs/Summary_2thematic_by_mask3.pmdl" –s –m
"N:/DATA/Outlines/Cnty_by_State/Cnty_NC_30m.img"
"N:/Acreage/NC12/Final/nc12sep_cdl_v1a_conf.img"
"N:/Acreage/NC12/Final/nc12sep_cdl_v1a.img"
"N:/Estiamtes/Acreage/WorkFiles_12_Sep/NC12/NC12Sep1_Cnf_CDL.mtx"
```

Output 1. Output from code to create batch file

The .pmdl model, referenced in the eighth line of Output 1, is run using ERDAS Imagine and goes through the following processes. First, any land cover categories with zero or null pixels in the CDL are removed to save on storage space. Second, the model assigns or recodes a unique code to every county and crop combination. For example, a cotton pixel (land cover category #2) located in county #145 would be assigned a value of 2145. Similarly, a soybean pixel (land cover category #5) located in county #6 would be assigned a value of 5006. There is an average of 62 counties per state and 130 possible different types of land cover within each state, therefore the table of pixel counts can be quite enormous, but much smaller than if no recoding were done. The final step of the model creates a matrix of pixel counts for each county and land cover combination where the classified pixels from the final CDL are the row values, the validation file pixels are the column values, and the pixel count for each scenario is the body of the matrix.

USING DATA STEP AND PROC DATASETS TO REFORMAT MATRIX TO COLUMN DATA

The resulting matrix file is imported into SAS using macros to rearrange the file into a column formatted dataset. The variables in the dataset are county id, validation land cover category, classification CDL land cover category, and the number of pixels. This data, calculated in ERDAS Imagine, is necessary to calculate the desired producer and user accuracies for each crop in each state and county. The producer accuracy, or omission error, is the probability that a validation pixel has been classified to the correct category in the CDL. An omission error occurs when a pixel from the CDL is excluded from the correct category. The user accuracy, or commission error, is the probability that a pixel from the CDL matches the validation pixel. A commission error occurs when a pixel from the CDL is included in the incorrect category.

```
DATA _NULL_;
    SET MtxColNames;
    LENGTH List Rename $2000;
    ARRAY Class(*) &ClassRng;
    DO i = 1 to HBOUND (Class);
      List = Trim(List)||' c'||PUT(Class[i],z3.);
      Rename = Trim(Rename)||' c'||PUT(i-1,z4.)||'c='||PUT(Class[i],z3.);
    END;
    CALL SYMPUT('ClassList',List);
    CALL SYMPUT('Rename', Rename);
RUN;

PROC DATASETS LIBRARY=Work NODETAILS NOLIST;
    MODIFY MtxMain;
    RENAME &Rename;
QUIT;
```

VISUALIZING COUNTY LEVEL ACCURACY ASSESSMENTS

To facilitate the ability of remote sensing analysts unfamiliar with SAS to investigate the quality and accuracy of each state's CDL, the data is visualized with a combination of device based graphics generating the maps and legends, and template based graphics to generate the county level horizontal bar and multiple scatter plots. These graphics are arranged in a web page framework using the html panel tagset. The resulting .pdf file can be seen in Figure 2.

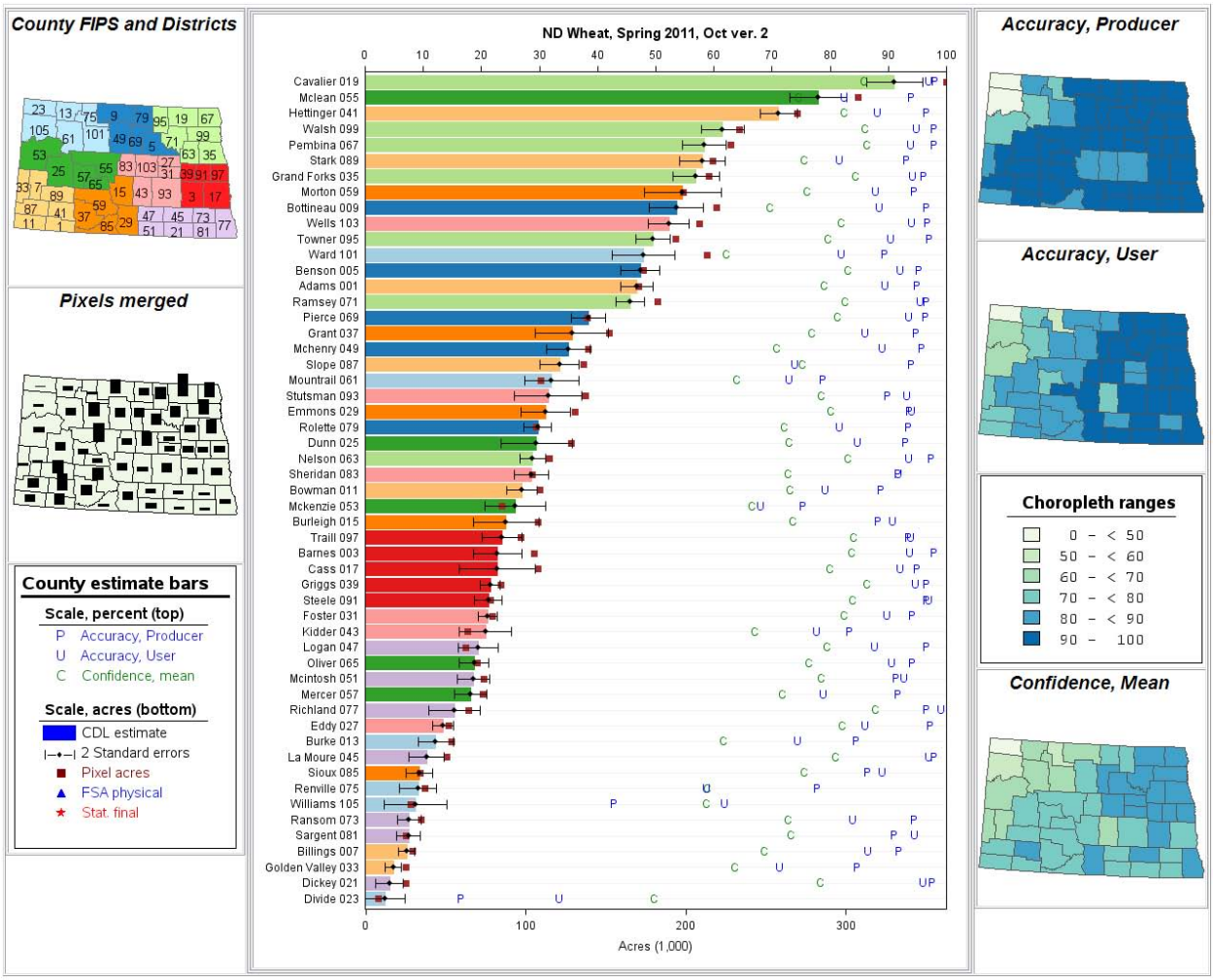

Figure 2. Final output showing statistics on spring wheat found in North Dakota in 2011. The data displayed in this figure is for demonstration purposes only and does not include any operational data.

Before detailing how the final product is constructed, it is useful to learn how to read the output. The dominant feature of the output is the bar chart in which each of the bars depicts the acreage estimation generated by regressing the CDL pixel count for a crop in a county to the acreage value determined by an independent USDA survey. The color of each of the individual bars corresponds with the color of the county in the choropleth map in the top left corner. Each group of counties within a state is considered an agricultural statistics district by the USDA. Matching the color of the bar to the color on the map aids the analysts when investigating a county on the chart to its geographical location within the state. Each bar has a two-sided whisker plot stretching 2 standard deviations from the acreage estimate calculated in the linear regression. The red points, found usually within close proximity of the top of the each bar, is the corresponding CDL crop pixel count in each county which was used as the dependent variable in the linear regression used to calculate the acreage estimate. Still within the bar chart, using the scale found on the top of the chart, are three values plotted on a line: P, U, and C. The letter P stands for the producer accuracy, the letter U stands for the user accuracy, and the C stands for the confidence value. The classification of pixels within the CDL is derived by an extremely complex decision tree. The confidence value is a measure of how complex the classification was for each pixel where values closest to 100 indicate a pixel that was easy to identify. The accuracies are calculated with Proc SQL and the confidence calculated with Proc MEANS. On the right side of the bar chart are three choropleth maps showing the corresponding producer accuracy, user accuracy, and confidence value for each county in the state. The final piece of the output is a map of the counties, found to the left center of the bar chart, titled “pixels merged” in which each county has an individual bar corresponding to the total crop pixel count where taller bars indicate more crop acreage.

Each of the pieces within this display are useful for the analysts who create the CDL. All of the values work in concert with one another to paint a picture of how accurately a particular crop is classified. Using the bar chart above an analyst is able to quickly identify the counties with the highest estimated acreage and their corresponding accuracies, In the graph above Cavalier county and McLean county have the highest estimated spring wheat acreage and their producer and user accuracies are greater than 80% in McLean county and closer to 100% in Cavalier county. Analysts can also use the output to identify areas that might need some improvement. Using the three choropleth maps on the right side of the output looking at the top left corner of the state has the lowest measured producer and user accuracies and also the lowest confidence values. Exploring the same area of the "Pixels Merged" chart on the left hand side of the output shows there are very few pixels acres of spring wheat in that part of the state. If that area is known to have more spring wheat acres an analyst may decide that it is necessary to find additional satellite imagery or survey data in that area before attempting to run the classification again.

BUILDING THE PIECES

The first step in constructing the output that will be viewed by the CDL analysts is defining a color range for the bar charts and the choropleth maps. This process is done using the widely accepted Brewer defined palettes which are generated into SAS datasets and macro variables using SAS programs developed by Michael Friendly. After the color palette and data ranges are chosen the choropleth ranges legend found on the upper right of the output is created before production of the output begins. The county estimate bar legend is also created pre-production using the annotate facility but with non-palette colors.

The bar chart is built using PROC TEMPLATE and numerous calls for a bar charts and scatter plots. It is necessary to define multiple x and y axis formats since half of the values in the chart correspond to acreage totals and the other half are measured in percentages. Below is the code used to set the format on the y axis and the two individual x axis's, one located on the bottom of the chart for the bars measuring the acreage estimates and the second x axis located on the top of the chart used to measure the producer and user accuracies as well as the calculated confidence values.

```
PROC TEMPLATE;
    DEFINE STATGRAPH Bar_Chart_county;
    DYNAMIC _yAxis "Category var" _xAxis "Continuous var";
    MVAR _Title "Title" _SF_Counts "StatFinal count";
    begingraph;
      entrytitle _Title;
      layout overlay /
        x2axisopts=(linearopts=(viewmin=0 viewmax=100 tickvaluesequence=(start=0
                    end=100 increment=10)) offsetmin=0 offsetmax=0 display=(tick
                    values line ticks))
        xaxisopts=(offsetmin=0 offsetmax=0 LABEL="Acres (1,000)")
        yaxisopts=(tickvalueattrs=(size=9pt) display=(tickvalues line));

        REFERENCELINE Y= _yAxis / datatransparency=0.9;
```

After the different axis's have been defined the code to call each of the charts beginning with the horizontal bar chart followed by the two scatter plots to show the CDL pixel count depicted as brown square and the corresponding whisker plots of the upper and lower bounds of two standard deviations away from the acreage estimate.

```
        barchart     x= _yAxis y=eval(asort(_xAxis, retain=all)) /name = "Estimate"
                     orient=horizontal display=(fill) datatransparency=0.0 group=asd
                     index=eval(INPUT(asd,3.)/10);

        scatterplot y=_yAxis x=Pixel_Acres / name='Pixel Acres' XAXIS=X
                     markerattrs=(SYMBOL=SquareFilled size=7 color=brown);

        scatterplot y=_yAxis x=Estimate / name=Est_Error XAXIS=X
                     MARKERATTRS=(SYMBOL=DiamondFilled size=6 color=black)
                     xERRORLOWER=eval(Estiamte-2*sqrt(Variance))
                     xERRORUPPER=eval(Estimate+2*sqrt(Variance))
                     ERRORBARATTRS=(Pattern=Solid color=black THICKNESS=1);

        scatterplot y=_yAxis x=Conf_Mean / name='Conf_Mean' XAXIS=X2
                     MARKERCHARACTER=eval(SUBSTR('C'||GeoLevel,1,1))
                     MARKERCHARACTERATTRS=(size=9 color=green);
```

```
        scatterplot y=_yAxis x=Conf_Mean / name='CorProd' XAXIS=X2
                    MARKERCHARACTER=eval(SUBSTR('P'||GeoLevel,1,1))
                    MARKERCHARACTERATTRS=(size=9 color=green);

        scatterplot y=_yAxis x=Conf_Mean / name='CorUser' XAXIS=X2
                    MARKERCHARACTER=eval(SUBSTR('U'||GeoLevel,1,1))
                    MARKERCHARACTERATTRS=(size=9 color=green);

ENDLAYOUT;
ENDGRAPH;
END;
RUN;
```

HTML PANEL TAGSET

With the bar chart and legends created, the next step in the process is to use the HTML Panel tagset with an ODS destination to combine the bar chart and legends with maps created using PROC GANNO. The HTML Panel will be defined to have 3 columns. The first column will contain a map of the state with colors corresponding to the bar chart, a map of the state with an individual bar in each county measuring crop pixel counts, and a legend for the bar chart. The second column will contain the extensive horizontal bar chart. The third column will contain two choropleth maps for the producer and user accuracies, a legend for the choropleth maps, and a map of confidence values for each county. Each piece within the column is defined as a specific cell within the code.

HTML Column 1

Below is the code used to define the HTML Panel with three columns and call the individual pieces of column one. Some of the detailed code used to create the individual maps has been removed to emphasize the process of creating the HTML panel tagset. The call for the ODS tagset begins the process of creating the annotated map where each agricultural statistics district and the counties within are given a specific color corresponding also with the bar chart. Next a map of the state is created using PROC GMAP where each county has a bar whose height corresponds to the total CDL pixel count within each state. The final piece of the first column in the output relates to the legend, stored as an image on the local hard drive, which was created manually and prior to any accuracy assessments being run.

```
filename htmlpath "C:\Temp";
ods listing close;
ods tagsets.htmlpanel nogtitle path=htmlPath
    file="Final.html" style=Paired
    options(panelborder='1' panelcolumns='4 3')

ods tagsets.htmlpanel event=row_panel(start);
ods tagsets.htmlpanel event=column_panel(start);

DATA Anno_Map_ASD …
TITLE1 "County FIPS and Districts";
PROC GANNO ANNOTATE=Anno_Map_ASD;

TITLE "Pixels merged";
PROC GMAP …

DATA My_anno…
goptions xpixel=240 ypixel=300;
PROC GANNO ANNOTATE=My_anno;
```

HTML Column 2

Calling the ODS tagsets again completes the first column of the output and begins the second column set to contain the detailed horizontal bar chart and the numerous scatter plots. Below is the code used set up and run the template for the bar chart.

```
ods tagsets.htmlpanel event=column_panel(finish);
ods tagsets.htmlpanel event=column_panel(start);

ods graphics / reset noborder width=750px height=1000px
    imagename="&File" imagefmt=png noscale;

PROC SGENDER DATA=RevGraph.ND12 TEMPLATE=Bar_Chart_county;
    DYNMAIC _yAxis="StAsdCty" _xAxis="Estimate";
    FORMAT Estimate Pixel_Acres Thous. StAsdCty $SAC_Fips.;
    WHERE GeoLevel='Cy' AND ssYYmmmv="&SSyyMMMv" and CatName04="Swht';
RUN;
```

HTML Column 3

As before the call of the ODS tagsets closes the second column and opens the third column in the document. The third column will contain two maps, both created using PROC GMAP, of the producer and user accuracy where darker colors within the map signify high producer and user accuracies followed by a legend of the colored values, also stored as a picture similar to the legend used in the first column of the output, and a final choropleth map depicting the confidence level measured in each county. Below is the code used to fill in the remaining pieces of the column 3 and to close the ODS listing.

```
ods tagsets.htmlpanel event=column_panel(finish);
ods tagsets.htmlpanel event=column_panel(start);

TITLE "Accuracy, Producer";
PROC GMAP …

TITLE "Accuracy, User";
PROC GMAP …

Data My_anno …
Goptions xpixels=240 ypixels=200;
TITLE;
PROC GANNO ANNOTATE=My_anno;

Title "Confidence, Mean";
PROC GMAP…

ods tagsets.htmlpanel event=column_panel(finish);
ods tagsets.htmlpanel event=row_panel(finish);
ods _all_ close;
ODS LISTING;
```

CONCLUSION

Since the inception of the Cropland Data Layer the evaluation of its accuracy, measured by producer accuracy, user accuracy, and confidence values for each crop, has been focused at the state level. It is crucial the major crops in each state are accurately classified in each CDL since the acreage estimates derived from it are used to generate the official acreage estimates published by the Agricultural Statistics Board. Until recently it was nearly impossible to investigate beyond the state level into the county level accuracies for each crop in a state and still have the time needed to make changes to the inputs and reprocess the CDL. Using the power of PROC DATASETS, PROC MEANS, PROC SQL, SAS tagsets, and SAS GRAPH the wealth of data displayed in each output is available quickly, displayed efficiently, and is very useful for analysts who need to investigate the accuracy measures. In areas where poor accuracy is found analysts are now better able to identify where more farmer reported data may be necessary or where there might be a need for more satellite imagery. Without this SAS process defined above it would be up to each analyst to run far too many manual processes on each state for each crop of interest which would prove detrimental to processing efforts for all the CDLs within the predetermined deadlines during a growing season.

REFERENCES

Friendly, Michael. “SAS Graphic Programs and Macros” *York University*. 2013. Available at www.datavis.ca/sasmac.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

Audra Zakzeski
USDA/NASS
3251 Old Lee Hwy Room 305
Fairfax, VA 22030
703-877-8000
audra_zakzeski@nass.usda.gov
www.nass.usda.gov/research/Cropland/SARS1a.htm

Robert Seffrin
USDA/NASS
3251 Old Lee Hwy Room 305
Fairfax, VA 22030
703-877-8000
robert_seffrin@nass.usda.gov
www.nass.usda.gov/research/Cropland/SARS1a.htm