Paper PO-18

Analysis and Visual Review of Error Matrices in SAS® Stat Studio

Robert Seffrin, USDA-NASS, Fairfax, VA

ABSTRACT

The United States Department of Agriculture's National Agricultural Statistics Service uses ground reference data and satellite imagery to create an annual Cropland Data Layer (CDL) land cover classification product. The CDL and survey data are used to estimate crop acreage. A review of the quality of the CDL begins with a cross tabulation of the CDL against a validation data set to produce an error matrix, also referred to as a confusion matrix or contingency table. The power of SAS/IML® Studio is employed to calculate the statistics using SAS/IML and IMLPlus and to generate insightful linked graphics and maps to review the quality of the land cover classification at the state and county level.

INTRODUCTION

The National Agricultural Statistics Service (NASS) conducts hundreds of surveys every year covering the breadth of US agriculture from aquaculture and horticulture to the more traditional crop and livestock farms. Most of these surveys are sampled from stratified lists of farm and ranch operators. Crop acreage surveys are conducted from these samples and crop area is estimated. Satellite imagery is the basis for a mostly independent non-survey crop acreage estimate.

Satellite imagery offers an alternative acreage estimator for large area crops like corn and soybeans providing complete coverage for a state much like a census. The input requirements are very different: cloud-free imagery (multi-date preferred), reference data for training, and specialized software for classification of raw data into land cover types. The result is the Cropland Data Layer (CDL) which is a raster file where each pixel is assigned to a ground cover type. Also like a census, errors of omission and commission can create bias for a particular cover type if crops were estimated with just a simple pixel count of the CDL raster product. A portion of the reference data is set aside as a validation data set which is used to create an error matrix, also called confusion matrix or contingency table. This paper presents useful tools available in SAS/IML Studio for summarizing and reviewing the categorical data in the error matrix.

DATA

The raw data for this analysis was generated in a Geographic Information System (GIS) software application by matching the validation raster layer with the 2009 North Dakota CDL raster layer and tabulating the count of pixels of each category in the CDL for each category in the validation layer. The result is an error matrix where each row is a category from the CDL and each column is a category from the validation data set as depicted in Table 1. The diagonal contains the counts where validation and CDL agree. To graphically use the error matrix in SAS/IML Studio it has been restructured into a frequency data set with a column for the CDL class names, a column for the reference (Ref) class names and a frequency column.

Table1. Upper left hand portion of the error matrix which has 46 rows and 46 columns.

		Reference data							
		Corn	Sorghum	Soybeans	Sunflower	Barley	Durham wheat	Spring wheat	Winter wheat
Cropland Data Layer	Corn	158,132	109	3,023	11,90	177	86	2,292	85
	Sorghum	20	346	10	25	2	0	2	0
	Soybeans	5,200	164	290,034	1,075	258	38	1,435	44
	Sunflower	639	73	738	61,643	44	29	849	8
	Barley	36	10	23	61	30,469	965	2,868	499
	Durham wheat	86	5	21	125	3,029	169,073	19,699	332
	Spring wheat	2,436	45	1,294	1,668	19,373	31,712	592,030	4,812
	Winter wheat	83	0	24	17	741	231	1,626	51,730

ANALYSIS AND REVIEW

A natural way to visualize the error matrix is a mosaic plot where the size of the rectangles are proportional to the cell counts. Figure 1 uses the MosaicPlot.CreateWithFreq() method to visualize the tabular version of the error matrix. The method sorts by variable values and automatically assigns colors. Any category with less than a definable threshold (set to 3 percent here) is placed in the "Other" category. To determine which cover combination is represented by a rectangle it may be necessary to select a rectangle and find the selected record in the linked data table. There are many keyboard shortcuts to modify the plot interactively such as changing the threshold for the "Others" category (Ctrl+number [0-9]), changing from percentages to frequencies ("P"), and removing a select rectangle from the plot and analysis ("E"). Since this CDL is relatively accurate (crop accuracies in mid 80 to mid 90 percentages), the good classification rectangles overwhelm the plot and interest may be more about the confusion between classes. These good agreements could be eliminated by selecting each good match rectangle and using the “E” shortcut or they could all be removed programmatically with this IMLPlus and IML code:

```
   dobj.GetVarData( {"Ref","CDL"}, Classes );
   ConfusDiagOff = LOC( Classes[,1] = Classes[,2]);
   dobj.IncludeInPlots(ConfusDiagOff, FALSE);
```

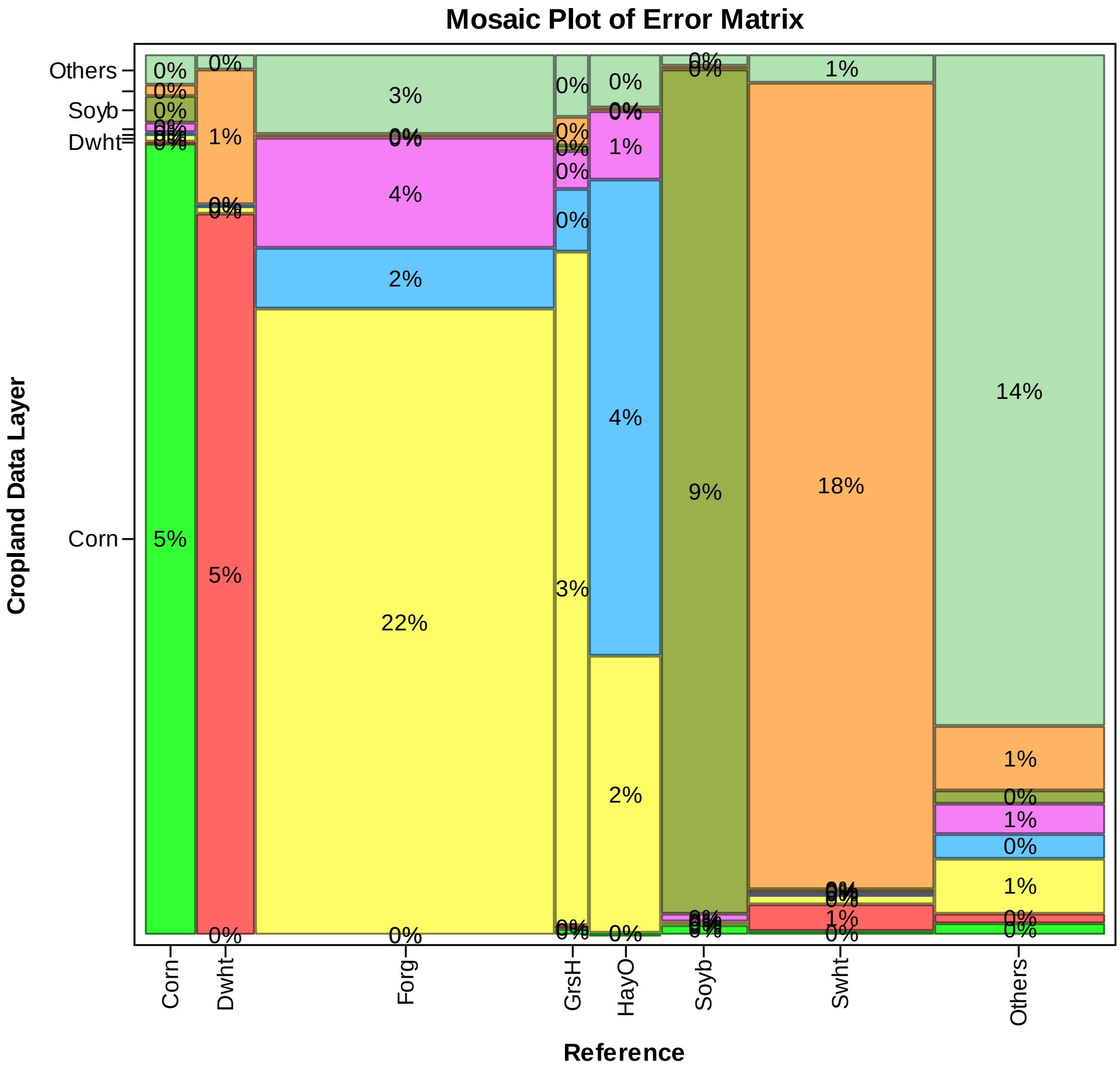

Figure 1. Mosaic plot of the error matrix with all combinations with abbreviated crop and cover names.

This plot is still overwhelmed by the forage=grassland herbaceous (Forg=GrsH) rectangle. These land covers are very similar to each other and not an important confusion relative to the main crops. If removed manually then Figure 2 focuses on the confusion between classes, more useful than the original mosaic plot in Figure 1 when reviewing the errors. In Figure 2 confusion with forage still dominates the plot but it is easier to see the confusion between two main crops of spring wheat (Swht) and durham wheat (Dwht) by the selected (diagonal hashed) rectangles which were practically invisible in Figure 1.

While the mosaic plot is useful to see the overall picture an accuracy measure provides a quantifiable number. The percent correct is the diagonal value divided by a marginal total. In remote sensing analysis the accuracy for the sum across reference columns is called the user accuracy while the sum across the rows of the target layer (here CDL) is called the producer accuracy. These are relatively easy to calculate using matrix algebra. This IMLPlus and IML code restructures the frequency table to recreate the error matrix where “dobj” below is the data object containing the frequency data:

```
   dobj.Sort({"nRef","nCDL"}, TRUE);
   dobj.GetVarData( "nRef", iU );
   dobj.GetVarData( "nCDL", iP );
```

```
dobj.GetVarData( "Correct", Freq  );
nClasses = UNIQUE(nRef);
Count    = NCOL(nClasses);
rFreq    = REPEAT(Freq, 1, Count);


/* Create error matrix based on User, apply Freq values */
nRefd      = DESIGN(nRef);
nCDLd      = DESIGN(nCDL);
nCDLdFreq = rFreq#DESIGN(nCDL);
eMatrix    = (rFreq#DESIGN(nCDL))`*nRefd;
```

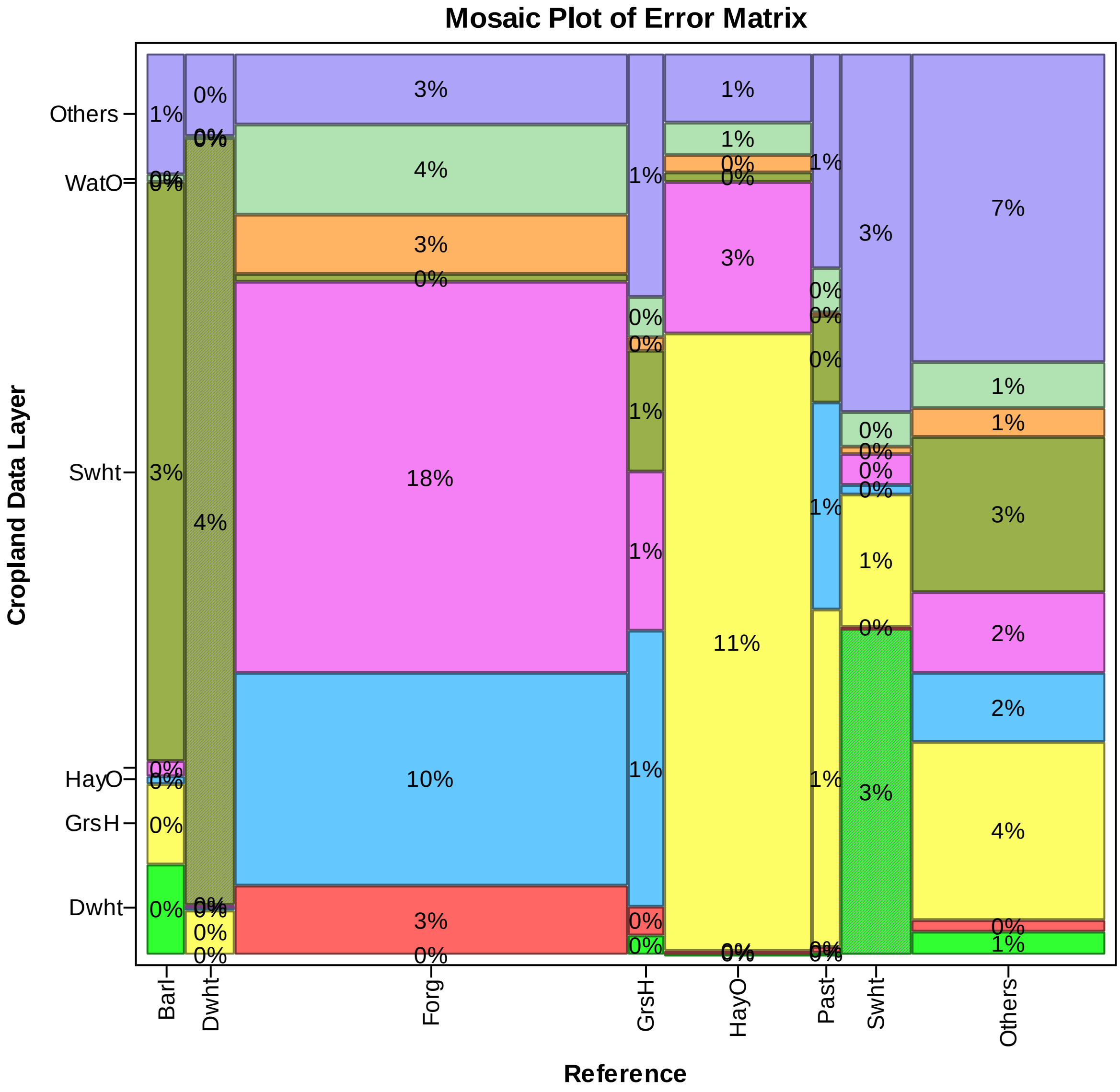

Figure 2. Mosaic plot with agreements and Forg=GrsH removed to focus on main crop confusion.

The accuracies of each class and their inverses, omission from the producer and commission from the user, as well as the conditional kappa may be calculated from each category with SAS/IML. Kappa is similar to chi-squared analysis and is a measure of how much better the classification is over chance agreement. A maximum likelihood estimate is used to calculate kappa from the perspective of the user and producer:

```
/* Preliminary calculations, marginal sums and products */
D            = VECDIAG(eMatrix);
SumRows      = eMatrix[,+];
SumCols      = eMatrix[+,];
MargProdSum = sumCols*SumRows;
SumRCprod    = SumRows#SumCols`;
N            = SUM(eMatrix);

/* Producer and user accuracies and omission and commission */
AccProd     = D/SumCols`;
Ommission   = 1-AccProd;
AccUser     = D/SumRows;
Commission = 1-AccUser;

/* Producer and user kappa */
KappaUser = (N#D- SumRCprod)/(N#SumRows - SumRCprod);
KappaProd = (N#D- SumRCprod)/(N#SumCols`- SumRCprod);
```

These statistics are only for error matrix diagonal cells and need to be realigned to match the dimension of the object frequency table which has all possible combinations with the following code.

```
AccClassAlln= {'Correct' 'TotProd' 'AccProd' 'Ommission'  'KapProd'
             'TotUser' 'AccUser' 'Commission' 'KapUser'};
AccClassAll = D ||SumCols`||AccProd ||Ommission || KappaUser
                ||SumRows ||AccUser ||Commission|| KappaProd;
dobj.GetVarData( "nRef", nRef );
dobj.GetVarData( "nCDL", nCDL );
nRefd = DESIGN(nRef);
nCDLd = DESIGN(nCDL);
/* expand the statistics to match the obj table */
eMatrixExp0 = nCDLd#nRefd*AccClassAll;
/* replace zeros with missing so zeros will not be plotted */
eMatrixExp = CHOOSE(eMatrixExp0=0,.,eMatrixExp0)
/* Attach to object table */
dobj.AddVars( AccClassAlln, eMatrixExp );
```

This allows the mosaic plot to be linked to plots generated from these added statistics such as the kappa plot in Figure 3. The blue reference line created with the DrawLine method is the equivalence between kappa for the user and producer. Selections in Figure 3 will also be highlighted in the data table and mosaic plot.

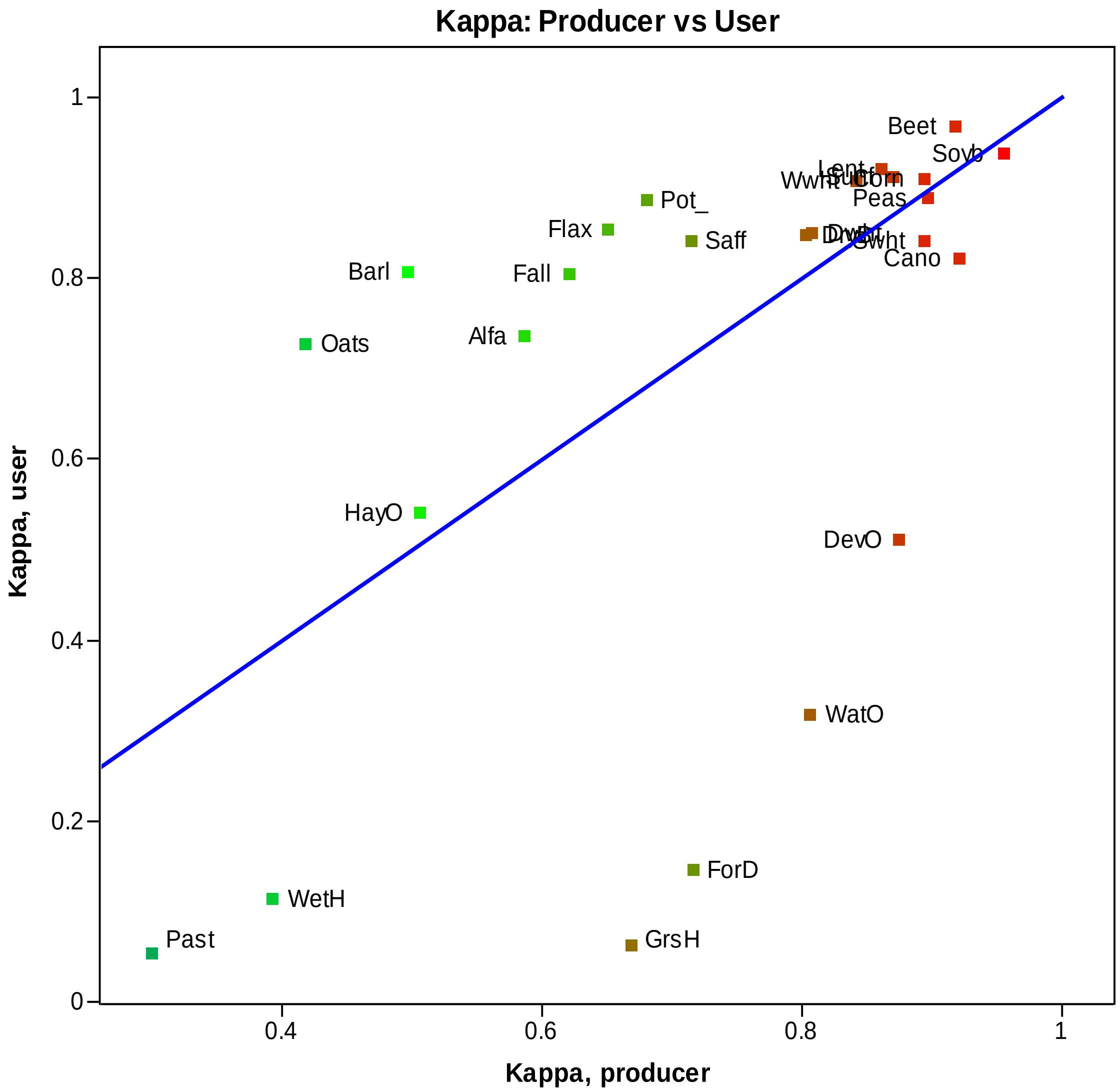

Figure 3. A scatter plot of producer kappa against the user kappa with color scaled across producer kappa.

If it is not clear whether the user or producer accuracy should be the reference then there is an option to normalize the error matrix to pre-defined marginal values. This also incorporates the off-diagonal values, including more information than the previous accuracy calculations and making the cell values comparable across different sample sizes. SAS/IML has the IPF (Iterative Proportional Fit) call. However the following simple module was written to normalize this matrix:

```
START MargFit_UniformMargs(InTable) GLOBAL(Iter, Dif);
    normTable = InTable;
    Count     = NROW(normTable); /* get size */
    MargTarg  = J(1,Count,1);    /* Fit all margin values to 1 */
    Dif       = 100;             /* set initial difference */
    Iter      = 0;               /* iteration counter, info only  */
    StopCrit  = 0.002;           /* max difference between target margin and sums */
    StopIter  = 100;             /* max number of iterations */
    /* check for zeros if any add .5 to all cell values*/
    IF ^ALL(normTable) THEN normTable = normTable+.5;
    /* Loop until criteria met */
    DO UNTIL (Dif < 0.002 | Iter > StopIter);
          t0r = normTable/normTable[,+]#MargTarg;/* adjust rows */
          t0c = t0r/t0r[+,]#MargTarg;            /* adjust columns */
          Dif = ABS(MargTarg`-t0c[,+])[+];       /* check differences */
              + ABS(MargTarg-t0c[+,])[+];
          normTable = t0c;                       /* update table and iteration */
          Iter      = Iter + 1;
    END;
    RETURN normTable;
FINISH;
```

Figure 4 is a plot of the producer accuracy against the normalized accuracy. The normalizing process tends to reduce the accuracy on the high end and increase the accuracy on the low end relative to the producer accuracy as more information is incorporated.

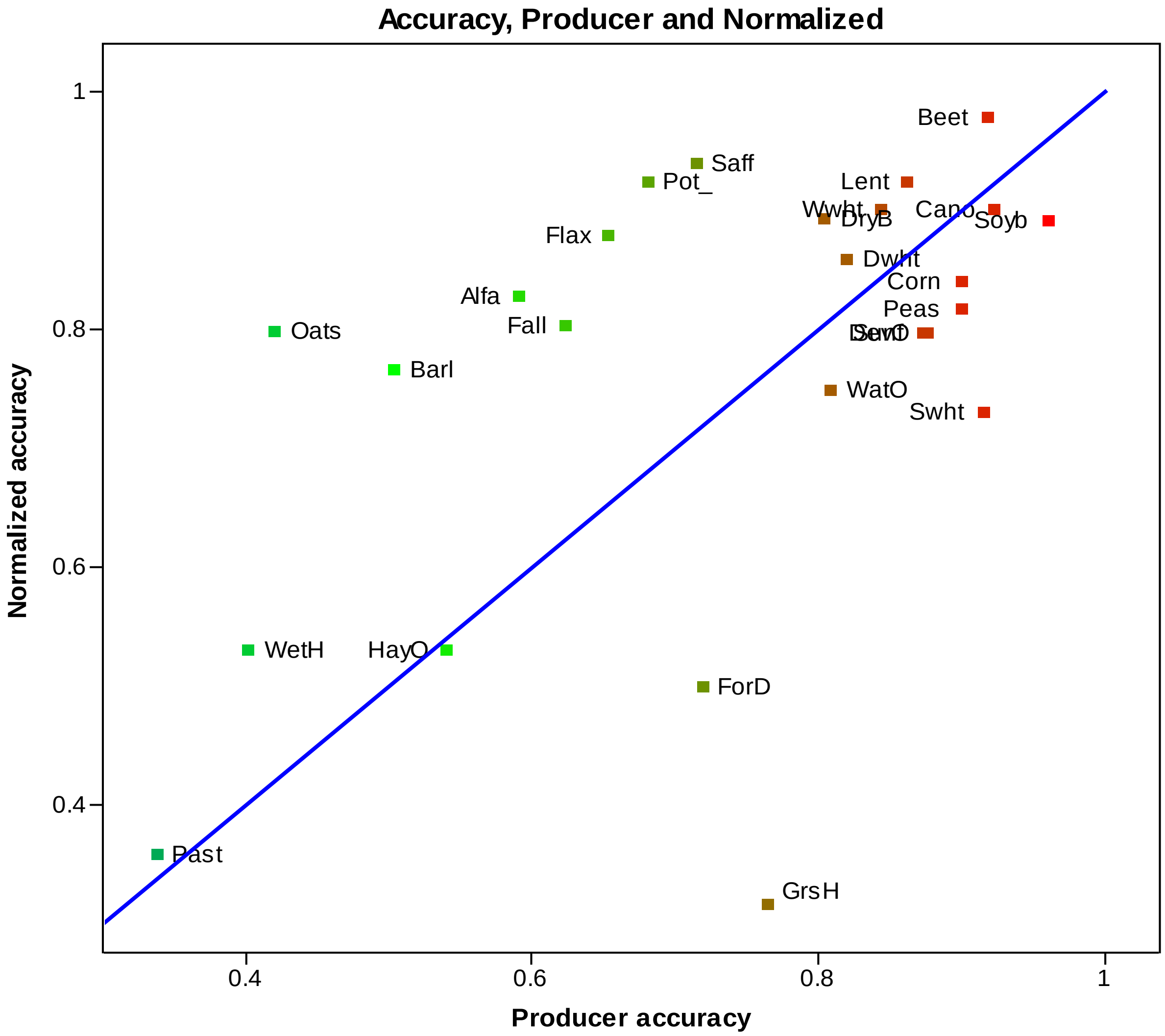

Figure 4. Scatter plot of producer accuracy against the normalized accuracy, color scaled across producer accuracy.

County Spatial Review

NASS also produces acreage estimates at the county level for major crops. This section reviews the quality of the CDL across counties for spring wheat, the largest crop in acreage and value in North Dakota. Error matrices and derived statistics were generated for each county of the North Dakota CDL. To display county boundaries on a map an ArcGIS shapefile was restructured into a polygon data format compatible with the PolygonPlot class. This is used purely as an annotation layer and not linked to the statistics table. The county crop statistics table includes the coordinates of the centroid of each county in the same map projection as the county boundaries.

Figure 5 shows the relationship between the user kappa and producer kappa for spring wheat across counties in a scatter plot. From the perspective of the map producer most counties have a kappa above 0.8. The kappas are more disperse from the perspective of user. Since this is spatial data it is useful to see if there are any patterns in the spatial distribution to locate the data strengths and weaknesses.

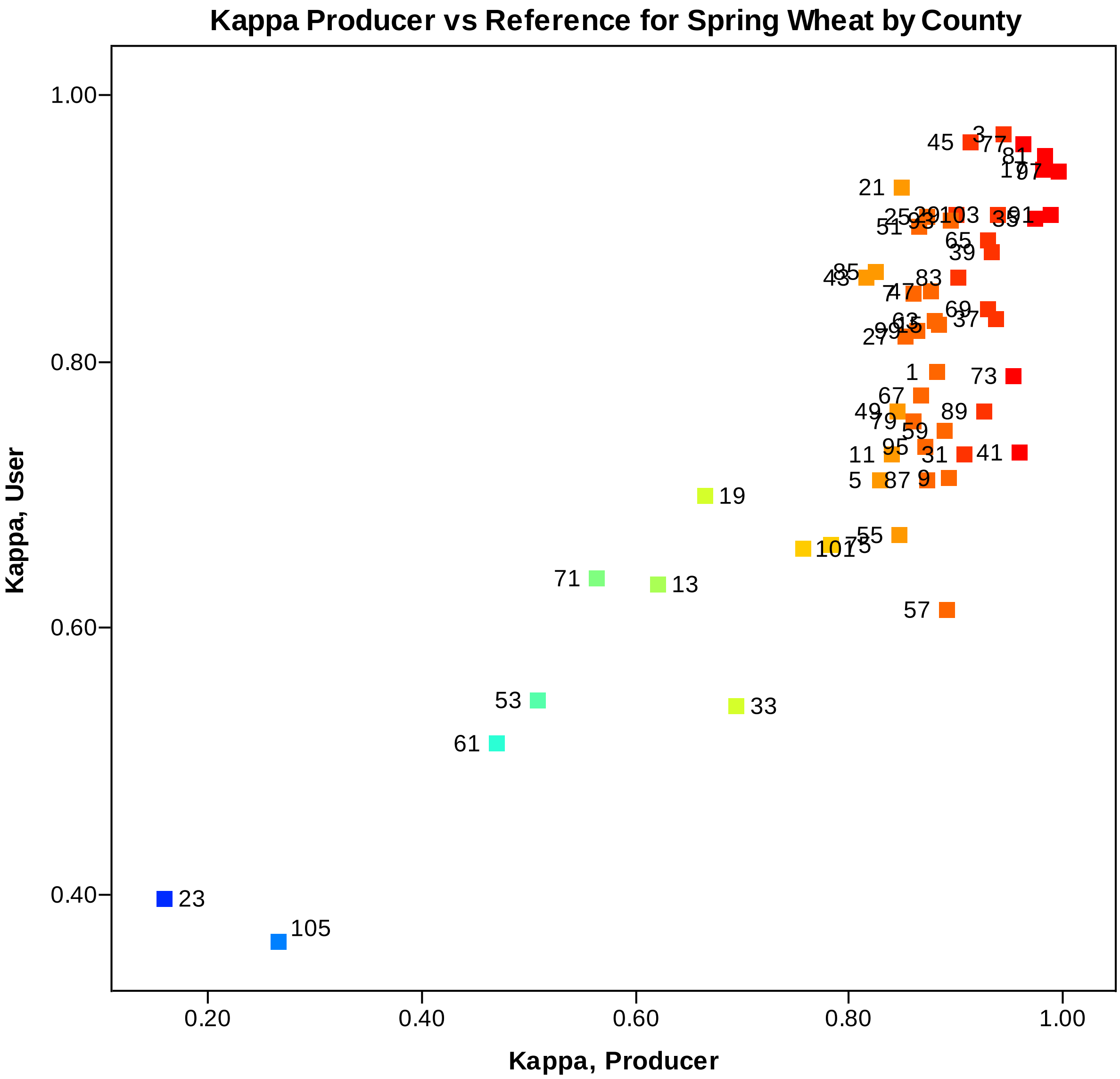

Figure 5. Distribution across counties of kappa for producer and user, color scales by increasing producer kappa value.

```
/* color code observations by values of KapProd */
objCtyCrp.GetVarData("KapProd", z );
SubDiv    = DO(0.1,1.0,.05); /* determines colors and values */
numColors = NCOL(SubDiv)-1;
Colors    = BLUE // CYAN // YELLOW // RED ;
ColorMap  = BlendColors( IntToRGB(Colors), numColors );
c         = RGBToInt( ColorMap );
DO i = 1 TO NCOL (SubDiv)-1;
    idx = LOC( z >= SubDiv[i] & z < SubDiv[i+1] );
    IF TYPE(idx)^='U' THEN
    objCtyCrp.SetMarkerColor( idx, c[i,] );
END;
```

The spatial distribution shown in Figure 6 was created with the ScatterPlot class and then underlain with the county boundaries using the SAS supplied DrawPolygonsByGroups module. Colors were defined in the data table and are the same and linked to the points in Figure 5.

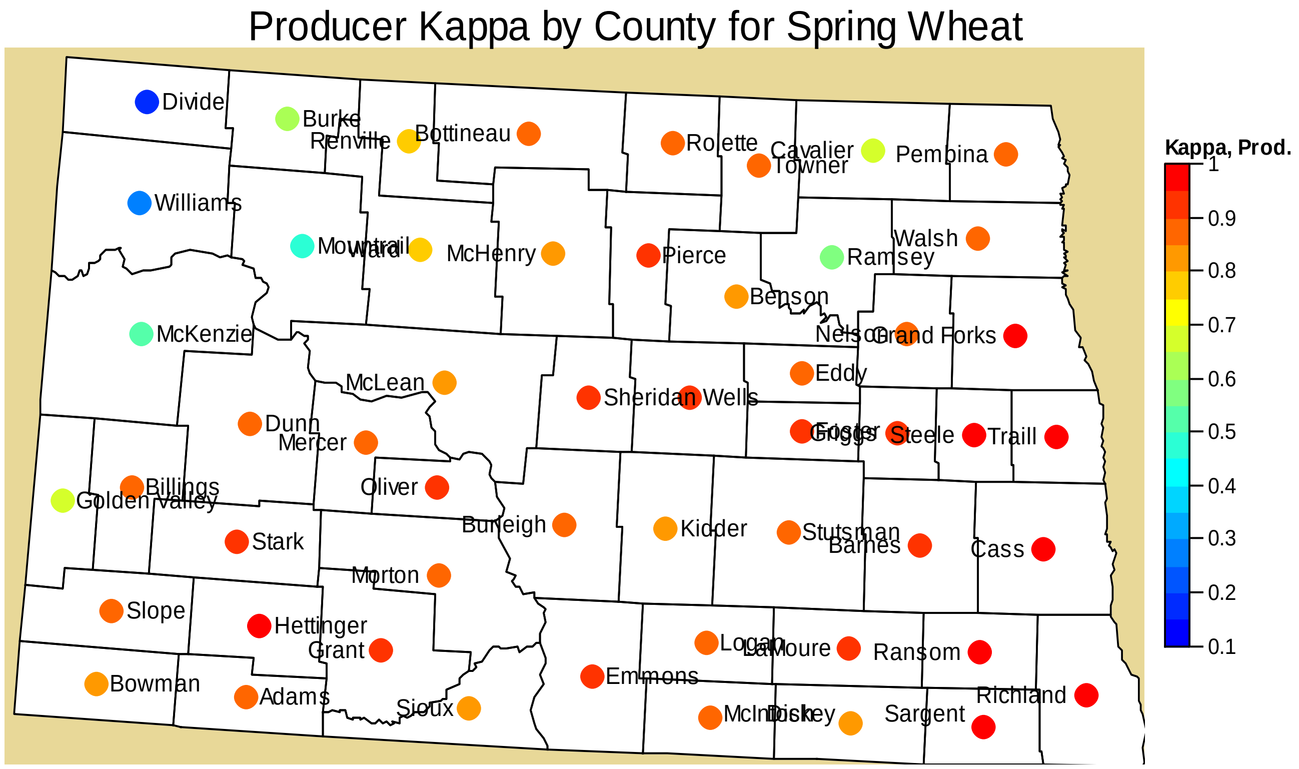

Figure 6. Distribution of producer kappa by county, using standard scatterplot options.

The legend in Figure 6 was created in a drawing block so that it could be removed, tweaked and replaced. The DrawContinuousLegend module is provided by SAS and is new in Stat Studio 9.2:

```
pScatCty1.DrawRemoveCommands("Legend");
pScatCty1.DrawBeginBlock( "Legend" );
    Ticks          = DO(0.1,1.0,0.1);
    TickPos        = (Ticks-0.1)/0.9;
    LegendSizeFrac = {0.08, 0.6};
    run DrawContinuousLegend( pScatCty1, Ticks, TickPos
           , "Kappa, Prod.", 8, ColorMap, LegendSizeFrac, "OR" );
pScatCty1.DrawEndBlock();
```

Clearly there is a pattern with the lowest kappas across the northwest and two counties in the northeast. The crucial piece of information still missing is the importance of the crop in each county based on the acreage. An analyst with limited time may not be concerned if a county's crop classification has a low kappa if the acreage is low relative to other counties. Figure 7 puts these kappas in perspective by plotting a rectangle that approximates the actual acreage of spring wheat in each county, again bringing across the color coding of the scatter plot in Figures 5 and 6. Rectangles of fixed width across counties were chosen over squares to improve the visual comparison of the relative acreages between counties. Counties with only circles are not published by NASS for this crop. While the rectangles themselves are not selectable and not linked to the data table, the county centroids underneath are linked and can be selected.

The following code builds the rectangles with colors from the data table, with details commented throughout:

```
/* Get observation numbers were not crop for excluding unwanted marker color obs. */
objCtyCrp.SelectObsWhere( "Prod04", WHERE_NE, Crop );
/* Save these observation to vector */
objCtyCrp.GetSelectedObsNumbers(mNOTSelectedObs);
/* Get fill color of observation */
objCtyCrp.GetMarkerFillColor( ColorsAll );
/* Remove observations not for this crop */
ColorsCrop = REMOVE(ColorsAll, mNOTSelectedObs)`;
```

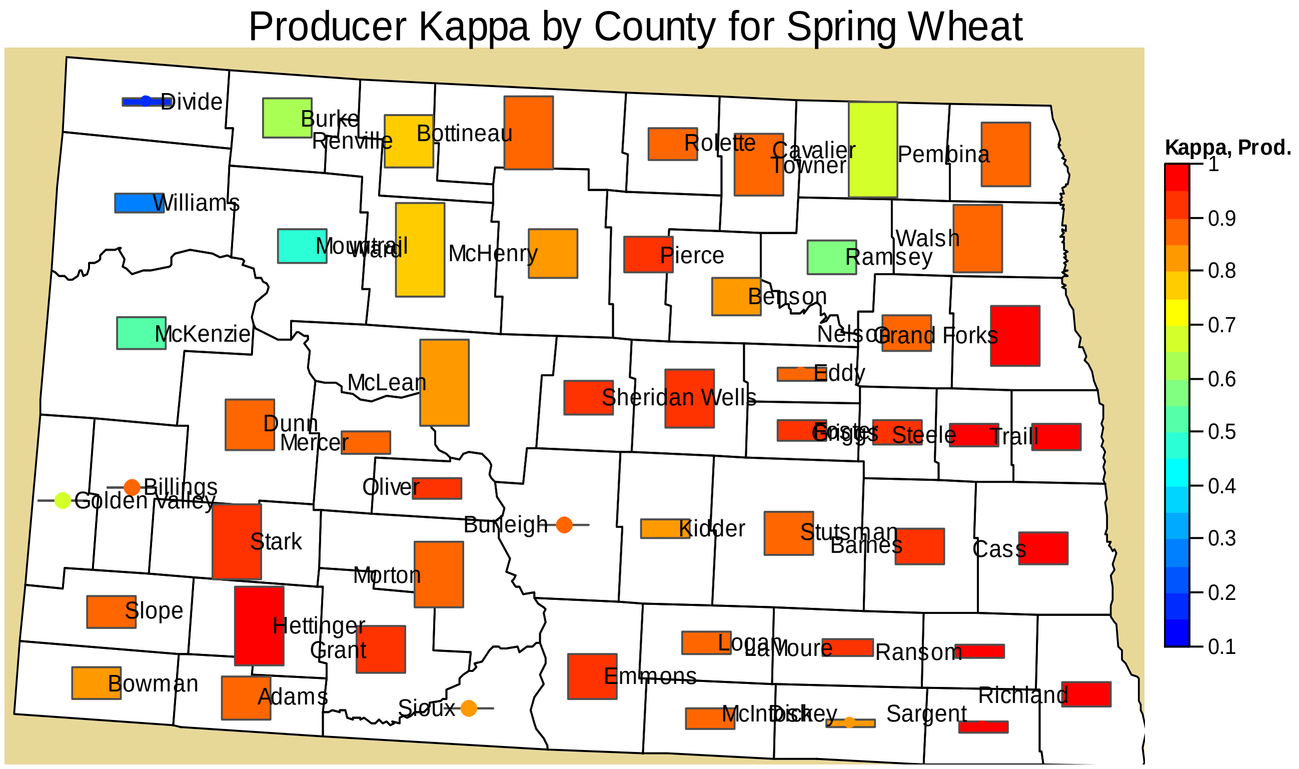

Figure 7. Rectangles show acreage of spring wheat in each county, color coded by producer kappa value from Figure 5.

```
/* Get the centroids and acreage data from data table */
objCtyCrp.GetVarData( {"Center_X","Center_Y","Acres"}, mSelectedObs, RectIN );
/* Concatenate to the color matrix */
RectINColor = RectIN||ColorsCrop;
/*Replace missing with zero so that counties without published acres still plotted */
RectINColor=CHOOSE(RectINColor=.,0,RectINColor);

/* Base is width in meters of the rectangles, arrived at by trial and error */
Base = 25000;
pScatCty1.DrawRemoveCommands("Squares");
/* Begin drawing block to create rectangles */
pScatCty1.DrawBeginBlock( "Squares" );
    pScatCty1.DrawPushState();
    pScatCty1.DrawResetState();
    pScatCty1.DrawSetRegion( PLOTBACKGROUND );
    pScatCty1.DrawSetPenColor( CHARCOAL );
    pScatCty1.DrawUseDataCoordinates();
    DO i = 1 TO NROW(RectINColor);
/*Acreage data is converted to meters and diagonal corners of rectangle calculated */
       DiagDist = RectINColor[i,3]*4046.86/Base/2;
       x1 = RectINColor[i,1]-Base/2;
       y1 = RectINColor[i,2]-DiagDist;
       x2 = RectINColor[i,1]+Base/2;
       y2 = RectINColor[i,2]+DiagDist;
       /* Color for this rectangle set */
       pScatCty1.DrawSetBrushColor( RectINColor[i,4] );
       /* Draw rectangle for this county */
       pScatCty1.DrawRectangle(x1, y1, x2, y2, TRUE );
    END;
    pScatCty1.DrawPopState();
pScatCty1.DrawEndBlock();
```

The above code was also run for the user kappa to produce Figure 8. The user kappas are generally lower than the producer kappas with the same general pattern of lower values in the northwest and northeast. Additionally, the user kappas are considerably lower than the producer kappas in the southwest. The low kappas in the northwest are due to confusion with durham wheat which is concentrate there. Other low kappa areas tend to follow the extent of herbaceous grassland which is the predominant cover in the western half of the state.

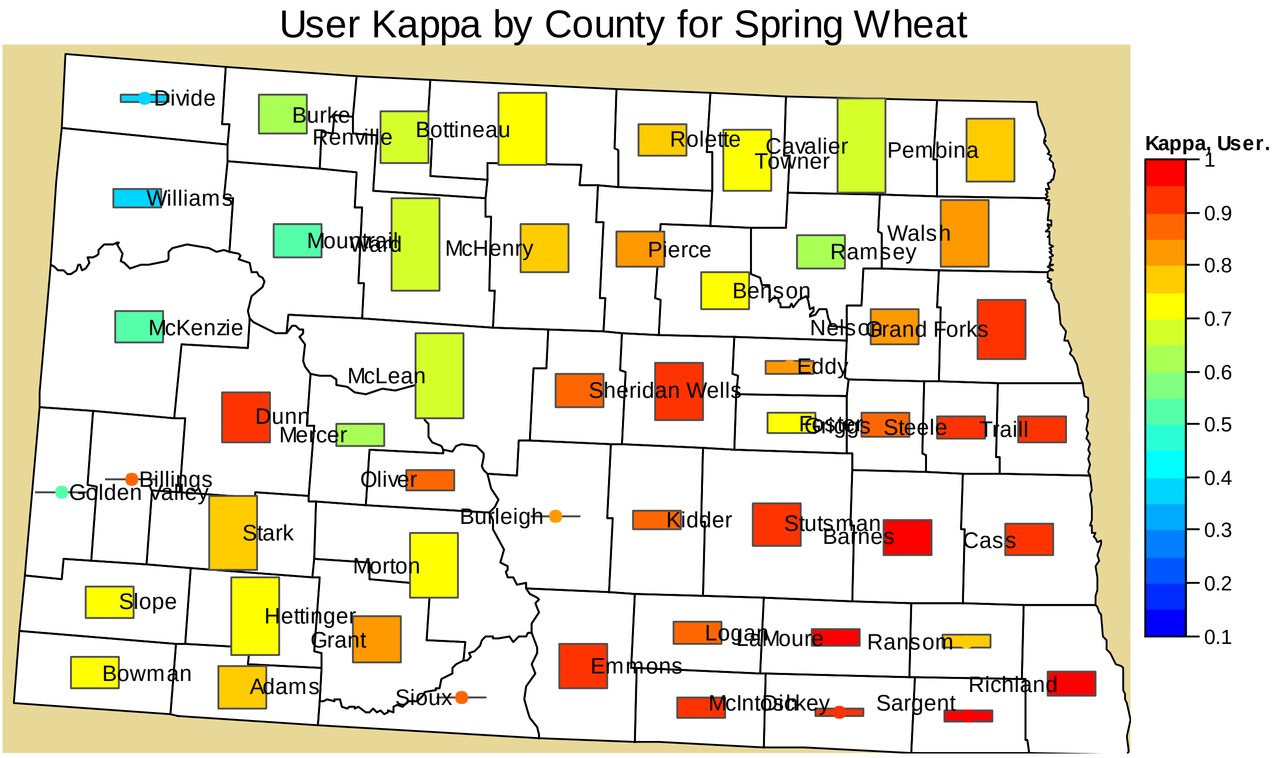

Figure 8. Rectangles show acreage of sprint wheat in each county, color coded by user kappa value.

CONCLUSION

Error matrix analysis has been around for some time but the SAS/IML and IMLPlus tools available through Stat Studio aid in the calculation and display of the data, especially large the tables from the cropland data layer. The mosaic plot summarizes the error matrix into one picture with interactive shortcuts to subset the classes of interest and manipulate the display. Accuracies and kappa values can be calculated with SAS/IML matrix functions and merged back to the data table and scatter plots created which are linked to other charts. County data may be mapped to show spatial trends and custom created layers such as the color coded acreage rectangles added to improve interpretation. Stat Studio is the ideal environment for exploring this type of data.

REFERENCES

Congalton, Russell G. and Green, K. *Assessing the Accuracy of Remotely Sensed Data—Principles and Practices* (Second edition), CRC Press, Taylor & Francis Group, Boca Raton, FL , 2009.

Foody, Giles M. Status of land cover classification accuracy assessment, Remote Sensing of Environment 80 (2002) 185– 201.

Friendly, Michael, *Visualizing Categorical Data*, Cary, NC: SAS Institute Inc., 2000.

National Agricultural Statistics Service (NASS) Online, 2009, http://www.nass.usda.gov/research/Cropland/SARS1a.htm> Accessed 2010, 01 June.

SAS Institute Inc. 2009. SAS Stat Studio 3.2 Help System. Cary, NC: SAS Institute Inc.

CONTACT INFORMATION

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

Robert Seffrin
USDA-NASS
3251 Old Lee Hwy
Fairfax, VA 22030-1504
Work Phone: 703-877-8000 ext. 155
Fax: 703-877-8044
E-mail: Robert_seffrin@nass.usda.gov

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are registered trademarks or trademarks of their respective companies.