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.

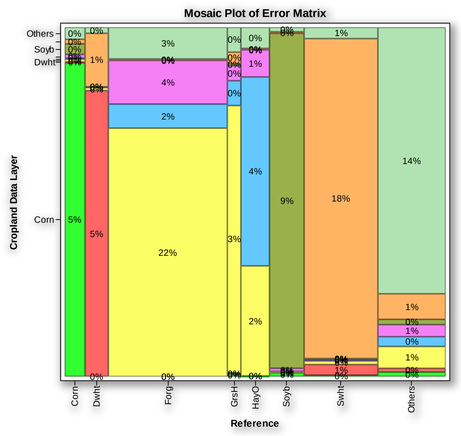
Satellite imagery offers an alternative acreage estimator for large area crops like corn and soybeans providing complete coverage for a state much like a census. The input requirements are very different: cloud-free imagery (multi-date preferred), reference data for training, and specialized software 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 analyse presents useful tools available in SAS/IML Studio for summarizing and reviewing the categorical data in the error matrix.

Data

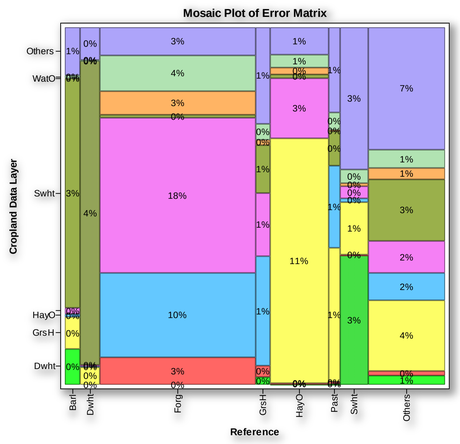
The raw data for this analysis was generated in a Geographic Information System (GIS) software application by matching the validation raster layer with the 2009 North Dakota CDL raster layer and tabulating the count of pixels of each category in the CDL for each category in the validation layer. The 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 the error matrix. The diagonal contains the counts where validation and CDL agree. To use the error matrix with graphic methods 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.

Mosaic Plot of the Error Matrix

A mosaic plot is a natural way to visualize the error matrix where the size of the rectangles are proportional to the cell counts. The figure below uses the MosaicPlot.CreateWithFreq() method to visualize the tabular version of the complete error matrix. Any category with less than 3 percent of the total is placed in the "Other" category.



The classification is relatively good and matches between the CDL and reference overwhelm the plot. To focus on the error part of the error matrix rectangles may be removed interactively or with IML plus code. The figure below has CDL=Reference values removed as well as forage=grassland removed to emphasis areas of confusion.



Error Matrix

		Reference data								
Cropland Data Layer		Corn	Sorghum	Soybeans	Sunflower	Barley	Durham wheat	Spring wheat	Winter wheat	continue →
	Corn	158,132	109	3,023	1,190	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	
	continue ↓									

Frequency Table

CDL	Reference	Frequency
Corn	Corn	158,132
Corn	Sorghum	109
Corn	Soybeans	3023
Corn	Sunflower	1,190
Corn	Barley	177
Corn	Durham Wheat	86
Corn	Spring Wheat	2,292
Corn	Winter Wheat	85
continue ↓		

Accuracies

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);
```

Normalization

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;
```

ScatterPlot class

After the accuracy statistics have been calculated (for values in diagonal of error matrix) they are expanded to match the row numbers of the frequency table and appended. The SAS supplied BlendColors module was used to assign colors across the producer kappa variable.

The ScatterPlot class was used to create the plot to the right to compare the kappa value between the producer and user. The DrawLine method to the Plot class was used to draw a blue equivalence line between the producer and user kappa.