Paper SD-09

Evaluating the Accuracy Assessment Methods of a Thematic Raster through SAS® Resampling Techniques and GTL Visualizations

Robert Seffrin, National Agricultural Statistics Service, US Dept. of Agriculture

ABSTRACT

The Cropland Data Layer (CDL) is a thematic raster layer of agricultural crops and other categories derived from satellite imagery and other data layers and trained on ground reference data. The CDL has been an annual product for the 48 contiguous states since 2008. The current accuracy assessment uses all data points/pixels from the interior of the validation fields. This approach introduces extensive spatial autocorrelation and ignores the 40 percent of the data points that fall in field edges. Field boundary pixels have been ignored in the past due to locational uncertainties. This project begins with the six million data points available for assessment from the 2012 Michigan CDL and applies resampling techniques from SAS®/SurveySelect procedure to address issues of varying field sizes, spatial autocorrelation and pixel location in the edge versus the interior of the fields. The results are summarized in customized Graphic Template Language (GTL) charts and alternatives to the current assessment methodology are discussed.

INTRODUCTION

This research is part of a project to develop a marginal estimator for crop acreage as an alternative to the CDL regression estimator that uses the June Area Survey (JAS) conducted by the National Agricultural Statistics Service. The regression estimator regresses the area of a crop from one to two square mile areas called segments from the JAS onto the same areas tabulated from the CDL, for a specified stratum and at the state level. This estimator has the same weaknesses of the JAS which are unquantified non-sampling errors and limited crop coverage in many county sized areas.

A marginal estimator would take advantage of the Farm Service Agency's Common Land Unit (CLU) which is the boundary for the smallest unit of land for reporting crops for administering crop programs. While the JAS samples one percent or less of crop land area, CLUs cover up to 95 percent of the crop land. The CLU data is split into a training portion for creating the CDL and a validation portion for testing the accuracy of the CDL classification. Our current validation data contains massive spatial autocorrelation and does not meet the statistical needs of a marginal estimator. This research will review the current validation data and ways to improve the error matrix that will provide the CDL production team and the public with a statistically sound accuracy measure, build confidence intervals around those estimates, and provide a basis for a marginal estimator.

BACKGROUND

CURRENT VALIDATION CONSTRUCTION

The validation data is built concurrently with the training data. When the crop data is imported and merged with the CLUs a subsample value from 1 to 10 is assigned to crop-CLU matches in the order of the records in the merge. Subsequent crop data updates and merges can create a new subsample number for each CLU. The general practice is that subsamples 1-7 are used for training and subsamples 8-10 are used for validation. Throughout the season as new crop data is retrieved and merged a particular CLU can alternate between training and validation.

Once the crop data is merged to the CLUs it is split into validation and training parts and converted to raster format. The validation is usually used in its entirety although for some states it has been subsampled. The validation raster is merged with the U.S. Geological Survey's National Land Cover Data (NLCD) raster to complete the training and validation data sets with non-crop categories. Both the training and validation CLUs are buffered inward by 1 pixel or 30 meters which removes 30 meters from outside edge and all remaining pixels are utilized. Figure 1 illustrates buffering and the terminology used in this paper. Removing the edge pixels reduces errors due to misalignment of spatial layers which can be up to 30 meters.

The error matrix is generated by overlaying the CDL with the validation layer and tabulating the categories of the overlaying pixels from each layer. The result is a matrix of counts whose row and column sums are in the margins and represent the errors of omission and commission. A marginal estimator uses this information to adjust for the biasedness of the CDL pixel count to estimate crop acreage. This assumes that the validation data set has no errors and has no spatial autocorrelation.

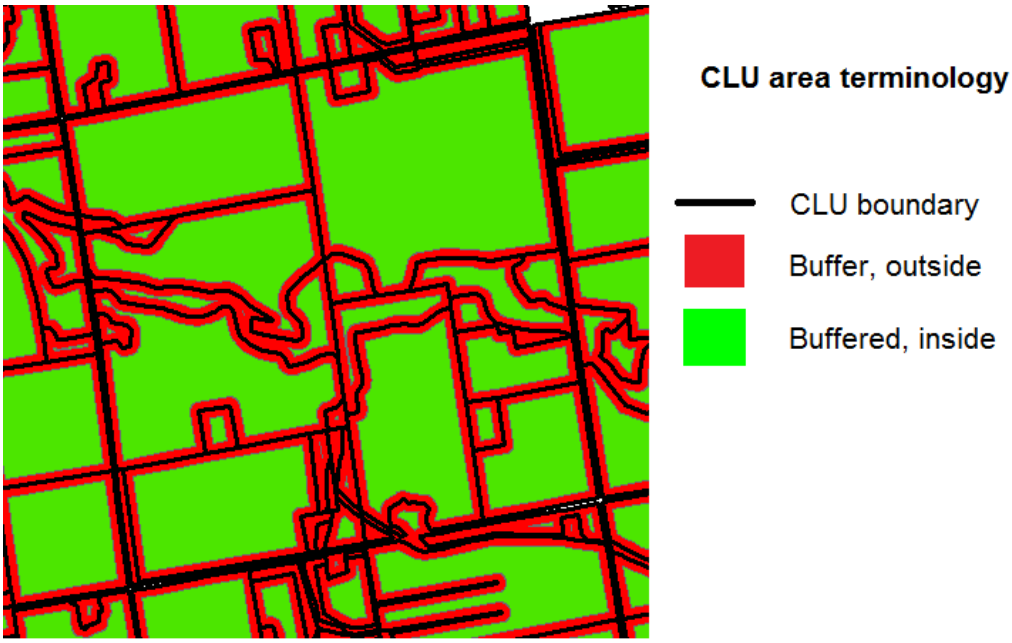

Figure 1. Terminolgy of Common Land Unit (CLU) and buffering.

SAMPLE DESIGN

For CDL production a minimum mapping unit (MMU) is not defined as prescribed for a formal classification scheme (*Congalton & Green*), although it is an option in a post production smart eliminate process (selective smoothing available in the NLCD tools for Erdas Imagine). All CDL production work is done at the pixel level. Pixels within a CLU have high sample dependence since they are a single crop and have a small contiguous extent. The size of buffered CLUs can range from one pixel to hundreds of pixels. Clearly the current method of using all or a large portion of pixels in a buffered CLU violates sample independence required for inferential statistics.

Congalton & Green argue against using single pixels because 1) pixel shape may not correspond to shape of cover, 2) positional accuracy for a single pixel is low and, 3) MMU is larger than a pixel. Buffering CLUs by one pixel eliminates issues 1 and 2, and the MMU is essentially one pixel in size. To avoid spatial autocorrelation only one sample should be selected from each CLU. This could be either a single pixel or a cluster of pixels treated as a single unit.

Much of the discussion of sample size in Congalton and Green assumes a cost constraint. With the extensive validation data set available here, the challenge to assess the CDL is less about cost and more about how to pick a sample that is independent yet represent the range of CLU sizes. For the marginal estimator Czaplewski and Catts recommend a sample of 500-1000 per area estimated. This would allow major crops to be estimated initially at the county level which are currently initially estimated at the state level.

EXPLORATION OF CLU-CROP DATA

The final 2012 Michigan CDL was used for this analysis. In order to get a complete picture of the crop areas, the CLU buffers were created by erasing the buffered CLU from the original projected CLUs using ESRI ArcGIS software. The buffered CLUs were then merged with the buffers to obtain complete state coverage of CLUs. The file was physically sorted to keep the buffer and buffered parts of CLUs adjacent in the attribute table.

A “Type” field was created to distinguish between sample versus validation CLUs and buffer versus buffered portions of the CLU:

Buffer or buffered area	Type =	Value range
Buffered (buf_type=2)	subsample number	1 - 10
Buffers (buf_type=1)	subsample number + 10	11 - 20
Buffered (buf_type=2)	CLU cover type (1 to 10) + 21	21 - 30
Buffers (buf_type=1)	CLU cover type (1 to 10) + 41	31 - 40

The buffer-buffered shapefile was rasterized on the value of row number plus 1. The Erdas Imagine pixel-to-table function was used to export all validation pixels and imported into SAS. The ArcGIS shapefile’s attribute table was imported into SAS to import the type value and other CLU attributes.

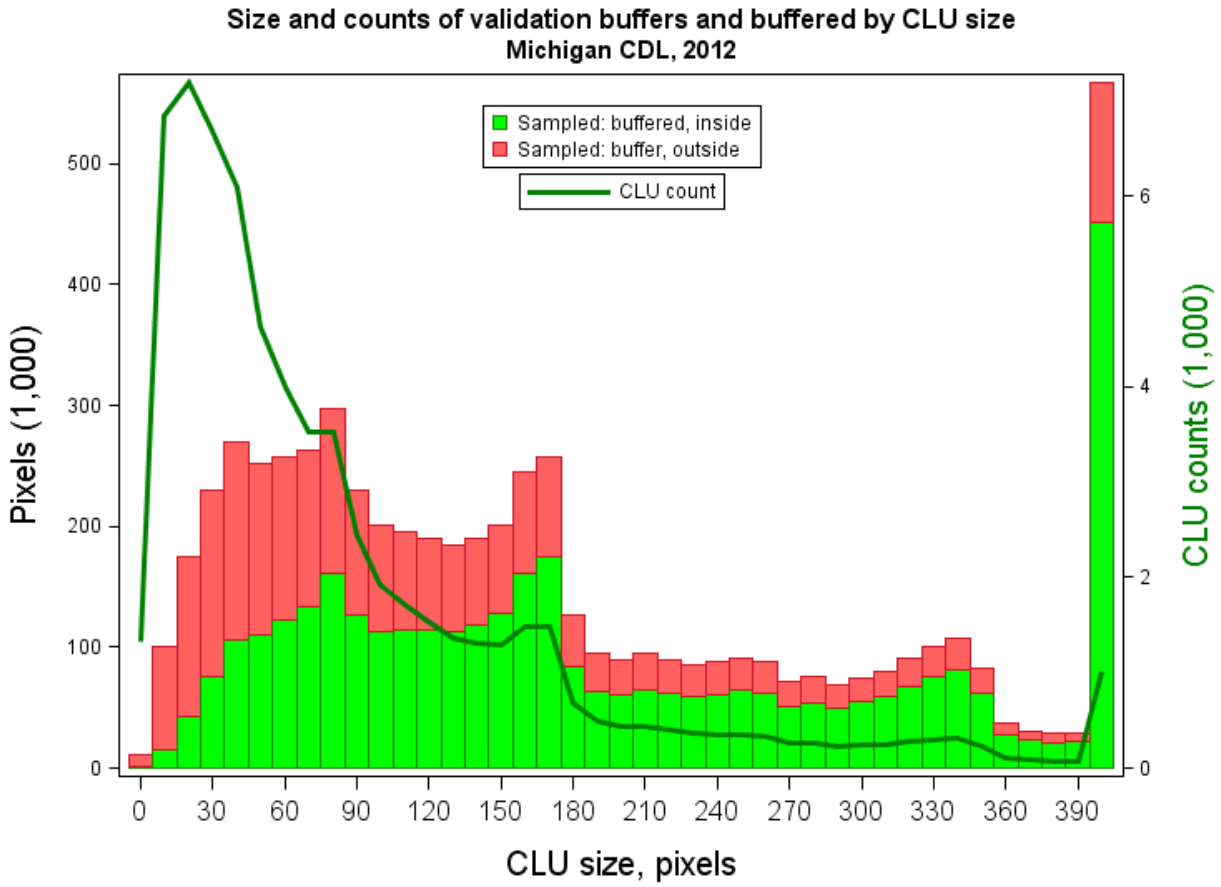

Figure 2. Distribution of validation CLUs by size and counts. All sizes greater than 390 pixels are in last bin.

Figure 2 shows the size distribution of the whole CLUs as the top of each histogram bar with the CLU buffers and the buffered areas proportions displayed in each bar. The line shows the number of CLUs in each histogram bin. Buffer areas account for 40% of all validation CLU areas while the buffered areas account for 60%. Ninety-two percent of all buffers are 80 pixels or less in size and become a smaller portion of CLUs as the CLU size increases. Current accuracy assessment uses buffered CLUs where there are a large number of small CLUs but they are a relatively small area of all buffered areas. The buffer area of a square CLU is larger than the buffered area when the side is less than 6.8 pixels long.

SAMPLING SCHEMES AND RESULTS

The current accuracy assessment methodology uses all or a large portion of the pixels in the buffered area of the validation CLUs. Not only does this create massive spatial autocorrelation, it is a source of optimistic bias (Verbyla) due to sampling from a homogeneous set of pixels. The current assessment also ignores 40 percent of the validation CLU area that is in the buffers. The justification to ignore the buffer pixels has been to discount the one pixel positional error of the input layers (CLU positional accuracy is much higher). This positional accuracy varies across the layers and is often less than half a pixel. The simplest sampling scheme to avoid spatial autocorrelation would be to randomly pick one pixel from each CLU. From figure 2 it is obvious that small CLUs would be disproportionately represented relative to the entire area of the CLUs and not representative the majority of the area. One way to avoid this is to sample proportional to size (PPS) and then treat multiple hits in a CLU as a cluster. This project will take a simpler approach and use individual pixels for analysis and ignore any remaining, although greatly reduced, spatial autocorrelation.

Bootstrap resampling was used for this analysis. In bootstrapping a sample of pixels with replacement is selected many times across the validation CLUs, and accuracies and variance calculated across these samples to estimate the mean and confidence intervals of the producer and user accuracy. The producer accuracy is the percent of pixels of a category in the validation map that matches the CDL while the user accuracy is the percent of pixels of a category in the CDL that matches the validation layer. The advantage of bootstrapping is that no assumptions are needed for the population parameters which are required for some estimators.

To explore the effect of CLU size on the accuracy estimate, 20 repetitions of a simple random sample of one pixel per

CLU was selected using the SAS SURVEYSELECT procedure.

```
SASFILE MI_CLUs_Valid_State LOAD; /* place data into memory */
OPTIONS nonotes;                  /* suppress NOTE that sample has only one choice */
PROC SURVEYSELECT
    DATA   = MI_CLUs_Valid_State
    METHOD = SRS              /* Simple Random Sample */
    N      = Temp_Hist_Bin    /* sample on average 1 pixel per CLU */
    SEED   = 9999
    OUT    = outFile
    REP    = 20               /* 20 repetitions */
    NOPRINT ;
  STRATA  buf_Type ;          /* buffer, buffered, buffer and buffered combined */
RUN;
OPTIONS Notes;
SASFILE MI_CLUs_Valid_State CLOSE ;     /* remove data from memory */
```

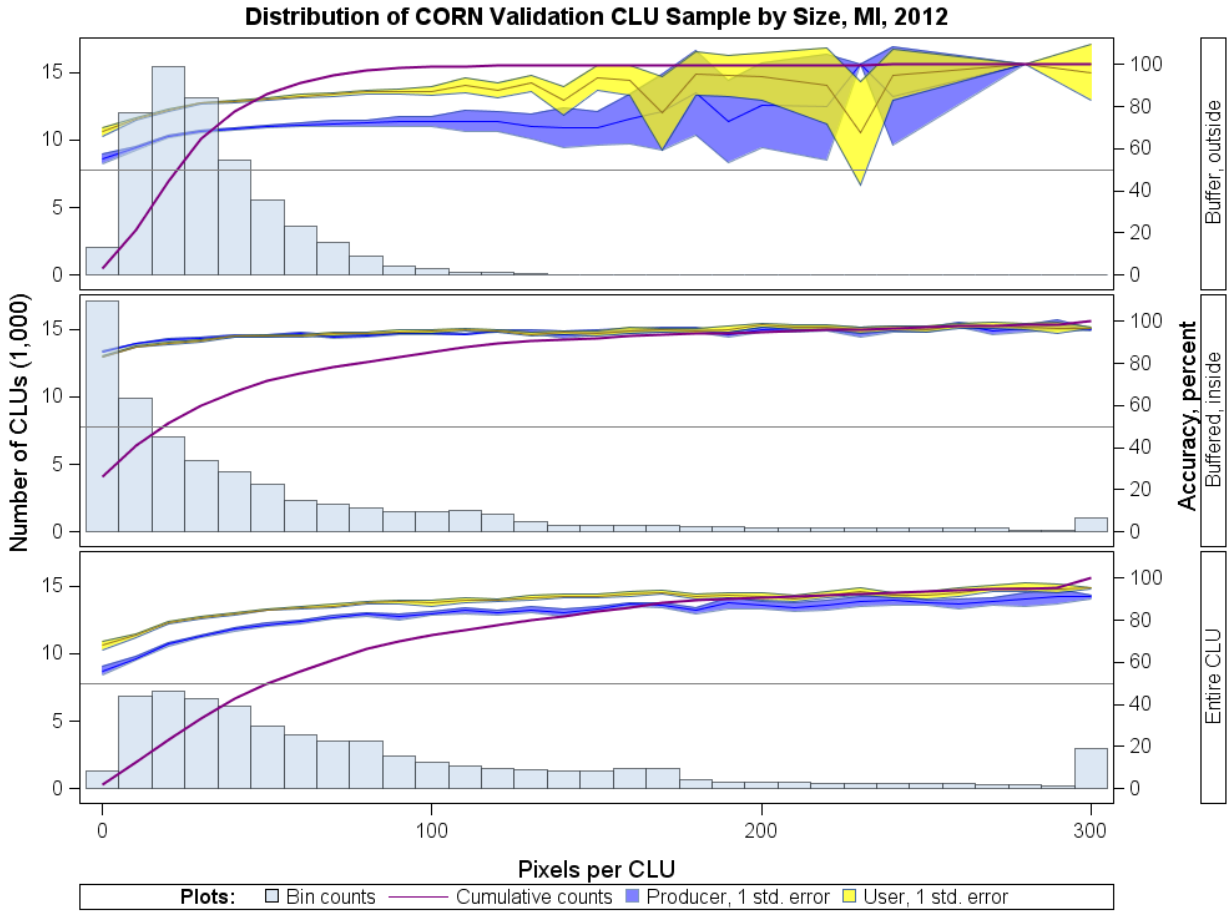

Figure 3. Sampling one pixel per CLU part or entire CLU, 20 repetitions, accuracy by CLU size.

The producer and user accuracy statistics are calculated in PROC SQL and the variances calculated in PROC MEANS. Figure 3 breaks out the selected CLUs by the parts of the CLU: buffer area, buffered area, and the entire CLU, by size along the x-axis, with the bands representing 1 standard error on each side of the accuracy. The errors for the buffers in the top plot increase dramtically because there are only a few CLUs with buffers that large. The middle plot is the buffered CLUs, which is currently used for CDL accuracy assessment where the producer and user accuracies for corn were 94.5 and 94.1 respectively. As expected, smaller CLUs have lower accuracy. The bottom

plot sampled from the entire CLU and has a wider error band than the buffered only area due to the inclusion of the buffers.

For CDL production it is not necessary to break out statistics by size but accuracy would be more representative of the CDL(and more useful to CDL users) if buffers, which account for 40% of the validation area were also represented. The next sampling scheme simulates a simple random sample on unbuffered validation CLUs by crop type based on the count of CLUs for each crop. This maintains a total sample size equivalent to one pixel per CLU and has the effect of sampling proportional to size (PPS) of the entire CLU. One hundred samples were created and summarized.

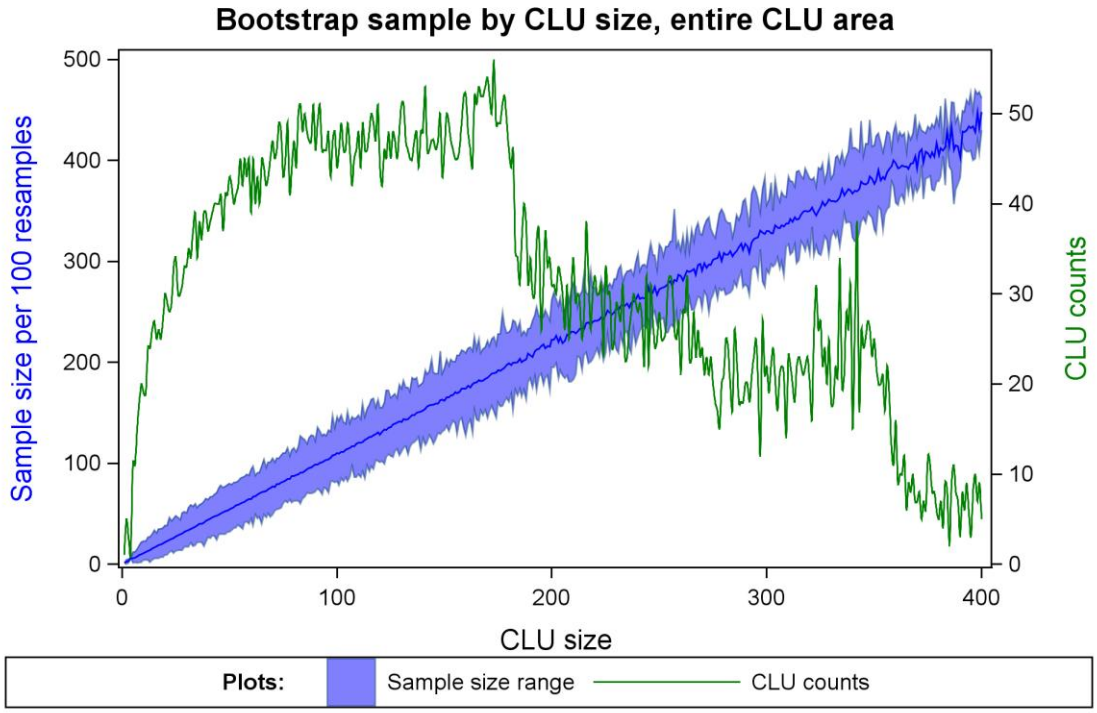

Figure 4. Results of SURVEYSELECT for 100 resamples across CLU sizes.

Figure 4 illustrates that PROC SURVEYSELECT (100 reps here) method of simple random sampling is essentially selecting samples proportional to size for this data set and the range of variation within each CLU size is somewhat consistent across CLU sizes.

Figure 5 summarizes both the current nonsampled accuracy assessment on the “All” lines and the result of the PPS sampling on the “Samp” lines. On each line is a summary by the parts of the CLU: buffer only, the buffered or inside area, and the entire CLU. The “All” estimates have no variance and are just points. The samples are summarized by the box plots with these parts: diamond is the mean, the vertical line near the center is the median, the box extends from 25th to 75th percentile while the lines extend to the maximum value within 1.5 times the interquartile range, and circles are outliers beyond that range.

For the producer accuracy the sampled and all pixel accuracies are in the same range because the random samples are drawn from the validation layer and the numerator and denominator of the formula both come from the validation layer. All of the sampled user accuracies are significantly lower than the “All” user accuracies as the number of highly correlated “good” pixels are reduced and/or more of the buffer pixels are included.

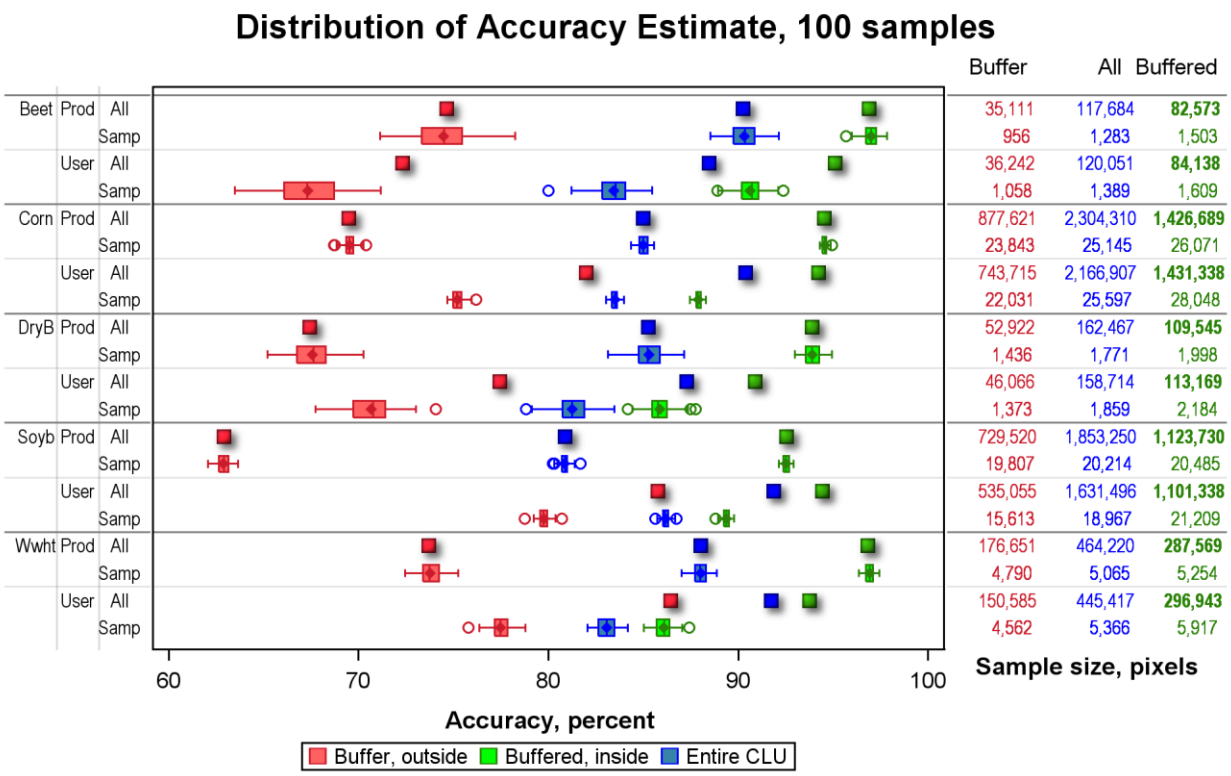

Figure 5. Accuracies of sampled and all pixels, by CLU part, crop, and user or producer.

CONCLUSIONS

A necessary and minimum requirement of accuracy assessment in the remote sensing community is a thorough documentation of the process used. The current procedure uses all pixels in the interior of the CLU and ignores the buffer. This results in spatial autocorrelation and leaves out about 40 percent of the crop area. A simple random sample across the unbuffered validation CLUs provides a compromise that represents some buffer area pixels but represents the larger buffered areas proportionately. This greatly reduces spatial autocorrelation while being relatively easy to incorporate into the production process. The resulting error matrix will allow a confidence interval calculation for the accuracy statistics and be a reasonable basis for a marginal estimator to calculate crop acreage independent of the JAS. A further refinement could be to stratify on CLU size to reduce the impact of greater variation of the smaller CLUs. The SAS SURVEYSELECT procedure greatly simplified the creation of data for this analysis. The Graph Template Language is a powerful and versatile tool to represent data and results.

REFERENCES

- Congalton, R. and K. Green. 2009. Assessing the Accuracy of Remotely Sensed Data: Principles and Practices. 2nd Edition. CRC/Taylor & Francis, Boca Raton, FL 183p. Congalton/Green book (C&G 2nd edition)
- Raymond L. Czaplewski, Glenn P. Catts, Calibration of remotely sensed proportion or area estimates for misclassification error, Remote Sensing of Environment, Volume 39, Issue 1, January 1992, Pages 29-43 (http://www.sciencedirect.com/science/article/pii/003442579290138A)
- SAS Institute Inc. 2012. SAS® 9.3 Graph Template Language: Reference. Third Edition. Cary, NC: SAS Institute Inc.
- USDA National Agricultural Statistics Service Cropland Data Layer. 2012. Published crop-specific data layer. Available at http://nassgeodata.gmu.edu/CropScape/. USDA-NASS, Washington, DC
- Verbyla, Dave. And Hammond, Tim, "How to Lie with an Error Matrix", Available at http://nrm.salrm.uaf.edu/~dverbyla/online/errormatrix.html [Accessed 1 March 2013]
- Wicklin, Rick, 2010, Statistical Programming with SAS/IML® Software. Cary, NC: SAS Institute Inc.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Graph templates available upon request. Contact the author at:

Name:	Robert Seffrin
Enterprise:	National Agricultural Statistics Service
Address:	3251 Old Lee Hwy
City, State ZIP:	Fairfax, VA 22030
Work Phone:	703-877-8000 ext. 155
Fax:	
E-mail:	robert.seffrin@nass.usda.gov
Web:	http://www.nass.usda.gov/Research_and_Science/index.asp