EVALUATING THE IMPACT OF TRAINING DATA PIXEL LEVEL BUFFERING ON AREA SAMPLING FRAME STRATIFICATION RESULTS AND CROP ESTIMATES

Claire G. Boryan, Zhengwei Yang, Robert Seffrin and Patrick Willis

USDA National Agricultural Statistics Service
Washington DC 20250, USA
Email: Claire.boryan@nass.usda.gov

ABSTRACT

Area Sampling Frames are used for surveys including crop acreage and yield, forests, and natural resource inventories and are the foundation of the statistical program of the USDA National Agricultural Statistics Service (NASS) and many statistical survey programs around the world. An automated area frame stratification method was recently implemented into NASS operations, which is based on the objective calculation of percent cultivation derived from the NASS geospatial Cropland Data Layers (CDLs). While autostratification consistently outperforms manual stratification in cultivated areas, we found that CDL-based pixel counting estimation consistently underestimated crop acreage. Previous research indicates that CDL classification accuracy is affected by training data pixel level buffering. We hypothesize that training data pixel level buffering will also affect the CDL based auto-stratification results and crop acreage estimation. This paper evaluates the impact of training data buffering on area frame stratification results and crop estimates. Preliminary results indicate that the crop acreage underestimation can be directly attributed to the training data pixel level buffering procedure.

Index Terms—Area sampling frame, automated stratification, buffered training pixels, crop estimates, land cover-based stratification, Cropland Data Layers

1. INTRODUCTION

Area sampling frames have been used in NASS since 1954 as a primary tool for conducting surveys to gather diverse agricultural information, notably planted acreage of major crops [1-4]. The NASS area frames are based on a stratification of U.S. land cover which classifies land into agricultural intensity groups (strata) based on percent cultivation of each primary sampling unit (PSU). The NASS Cropland Data Layers (CDLs) which are 30 m to 56 m crop specific land cover classifications created annually using satellite data are a relatively new data source for area frame stratification [5]. A new automated stratification method, which is based on the objective calculation of percent cultivation, at the primary sampling unit level, obtained from the CDLs, was recently integrated with manual review and editing for NASS operations [6-7].

While automated stratification consistently outperforms manual stratification in cultivated areas, both in research and production environments, one limitation of the automated stratification method is that the CDLs, as they are currently processed, consistently underestimate crop acreage [8]. The authors hypothesize that this underestimation is due to a processing procedure used in the preparation of the Farm Service Agency (FSA) Common Land Unit (CLU) training data [9]. During CDL production, the FSA CLU shape files are buffered inward one pixel (30 m) resulting in the exclusion of field edges, as well as small and/or narrow fields for use in CDL training. Previous research results indicate that training data buffering resulted in lower cultivation accuracies [10]. Further research on the impact of training data pixel level buffering on area frame stratification needs to be conducted to determine if there are significant differences in the percent cultivation of PSU computed from CDLs and in the obtained crop estimates.

This paper presents an assessment that evaluates the impact of buffering the FSA CLU training data, for CDL production, on area frame stratification results and crop estimates. In this study, Nebraska (NE), U.S. (Fig. 1) is selected as the study area because it is an important agricultural state in the U.S. and is a good example of the range of crops grown. The geospatial datasets used include: the NASS 2015 Nebraska area frame PSU dataset and two 2015 Nebraska CDLs created using USDA Nebraska FSA CLU training data preprocessed with 1) a 30m buffer or 2) no buffer.

The impact of training data pixel level buffering is evaluated based on 1) the percent difference in Nebraska area frame stratum acreage and total number of stratum PSUs and 2) a comparison of 2015 Nebraska state level corn, soybean and winter wheat planted acreage estimates obtained from 270 samples per stratification (30 m buffer vs. no buffer).

Fig. 1. Nebraska (NE) U.S. - Study Area for Area Frame Stratification results and Crop Estimate Comparison

2. DATA AND SCOPE

2.1. NASS Area Sampling Frames

NASS's primary area frame based survey is the June Area Survey (JAS) in which approximately 11,000 one square mile sample segments (land parcels) are visited by survey enumerators at the beginning of each growing season to collect crop type and acreage information. Estimates of crop acreage and livestock inventories are based on the JAS data. The accuracy of NASS survey statistics depends on the quality of the NASS area frames and the techniques used in their construction.

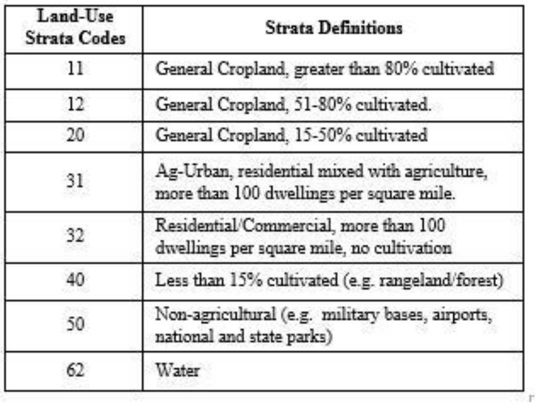
The NASS area frames are made up of stratified parcels of land, known as primary sampling units (PSUs), which are digitized to physical boundaries (roads, railroads, and rivers) on the ground. The NASS area frame stratification is based on percent cultivation of the land cover within PSUs. Table 1 illustrates NASS Nebraska land-use stratification codes and definitions. Once stratum definitions are assigned, all land is subdivided into PSUs, which are categorized into different strata. Selected PSUs are further subdivided into segments or sample units, and a segment is randomly selected from each selected PSU for enumeration [1].

2.2. NASS Cropland Data Layers (CDLs)

The CDL is an annual crop-specific land cover classification covering the continental U.S. [5]. The 2015 CDLs were created using a decision tree classifier Rulequest See5.0 software. ERDAS Imagine software is used in the pre- and post- processing of all raster-based data.

ESRI ArcGIS is used to prepare the vector-based training and validation data. Agricultural training and validation data are derived from the Farm Service Agency (FSA) Common Land Unit (CLU) Program. The United States Geological Survey National Land Cover Database 2011 was used as non-agricultural training and validation data for the 2015 CDLs.

TABLE 1. Land-Use Stratification Codes and Definitions Represented in the NASS Nebraska Area Sampling Frame



Land-Use Strata Codes	Strata Definitions
11	General Cropland, greater than 80% cultivated
12	General Cropland, 51-80% cultivated.
20	General Cropland, 15-50% cultivated
31	Ag-Urban, residential mixed with agriculture, more than 100 dwellings per square mile.
32	Residential/Commercial, more than 100 dwellings per square mile, no cultivation
40	Less than 15% cultivated (e.g. rangeland/forest)
50	Non-agricultural (e.g. military bases, airports, national and state parks)
62	Water

3. METHODOLOGY

3.1. Training Data Processing and Area Frame Stratification

This study evaluates two scenarios comparing the use of buffered and non-buffered training data to generate land cover classifications for automated area frame stratification and crop estimation. Nebraska 2015 is used to test the difference between two differently buffered versions of the CDLs. Both CDLs use the standard CDL processing methodology, identical satellite imagery and ancillary data inputs, training sample sizes and classification parameters. The only difference between the two Nebraska 2015 CDLs is the way the FSA CLU data are pre-processed (30 m buffer vs. no buffer).

To create the buffered training data, FSA CLU polygons (in shape file format) are buffered inward 30 m with the intention of eliminating mixed spectral training at field edges. The FSA CLU data are left in their original form for the non-buffered training. Both buffered and non-buffered CLU polygons are linked to corresponding FSA Form 578 database files, dated Oct 1, 2015, which contain crop-specific field-level information.

The two 2015 Nebraska CDLs are both recoded into two bit cultivated/non-cultivated data layers. The ESRI ArcGIS Zonal Statistics tool is used to stratify the NASS Nebraska area frame, based on percent cultivated cropland calculated within a PSU for each different buffer-type CDL, resulting in two Nebraska area frame stratifications for evaluation and to conduct the crop estimate comparison.

3.2. Estimation

A stratum level sample allocation was performed using each buffer type stratified area frame and the past five years of NASS JAS state level standard errors for each crop and stratum. Since the most recent year's standard errors best reflect current conditions, sample allocations were weighted for the years 2011 to 2015 respectively: 0.1, 0.1, 0.15, 0.25, and 0.40. The weighted allocations were then prorated across strata to arrive at the 2015 state total allocation of 451 sample units. There was a final manual adjustment to the number of repetitions and substrata for each stratum so that the total sample size would be the same for the two area frame stratifications (30 m buffer vs. no buffer). The allocation formula is the same used operationally for the NASS JAS and described by Reference [1].

For these two buffer type area frame stratifications, a random selection of PSUs within strata was run 270 times with a unique random seed for each run, resulting in 270 distinct samples for each of the two area frame stratifications. The selected PSUs, in the buffered area frame stratification, were merged with the official 2015 Nebraska CDL crop tabulations for the randomly selected PSUs (NASS, 2016). The selected PSUs in the non-buffered area frame stratification were merged with the unbuffered 2015 Nebraska CDL. Each PSU level crop tabulation was expanded using the PSU expansion value based on the stratum populations. The expanded PSU crop totals were summed to arrive at planted acreage estimates for Nebraska 2015 corn, soybeans and winter wheat at the state level. A t-test was conducted to test the difference in the mean of these sample distributions for each of the crop estimates.

4. RESULTS AND DISCUSSION

Figure 2 illustrates zooms of Nebraska 2015 CDLs created using training data with 30 m buffer (right) vs no buffer (left). The zero buffer and 30 m buffer FSA CLU shape files are overlaid on the resulting CDLs. Results, in Tables 2 – 4 show that, when using CDLs generated with the 30 m buffer vs. no buffer training data, there was a significant impact on the area frame stratification results and 2015 NE corn, soybean and winter wheat planted acreage estimates.

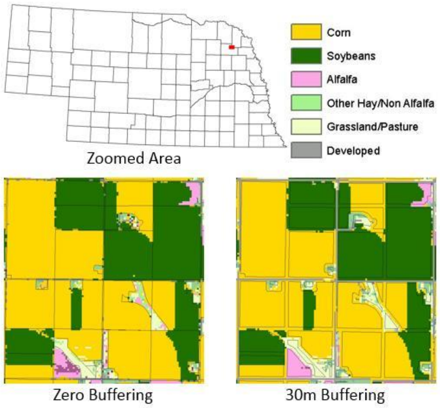
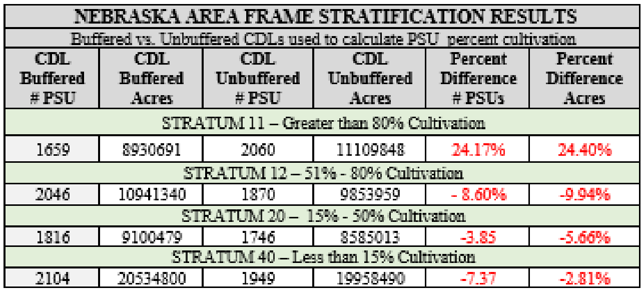


Fig. 2. Zooms of Nebraska CDLs created using training data with 30 m (right) vs no buffer (left). The Zero buffer and 30 m buffer FSA CLU shape files are overlaid on the resulting CDLs. The red dot identifies the location of the zoomed area.

4.1. Stratification Results

Table 2 illustrates the stratification results that show that when no buffer was applied to the training data and the resulting CDL was used to stratify the Nebraska area frame PSUs, 24.40% more cropland acres were identified in the highly cultivated strata (stratum 11) and reduced crop acres were identified in the lower cultivation strata (strata 12, 20 and 40), which would likely impact crop estimates.

TABLE 2. Stratification Results



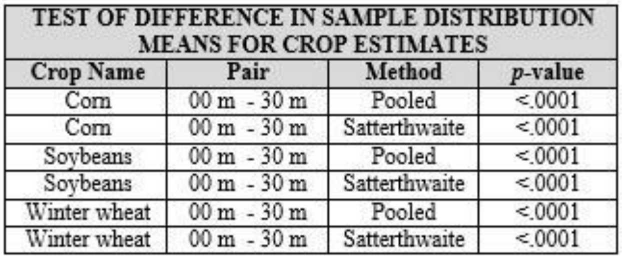
NEBRASKA AREA FRAME STRATIFICATION RESULTS					
Buffered vs. Unbuffered CDLs used to calculate PSU percent cultivation					
CDL Buffered # PSU	**CDL Buffered Acres**	**CDL Unbuffered # PSU**	**CDL Unbuffered Acres**	**Percent Difference # PSUs**	**Percent Difference Acres**
STRATUM 11 – Greater than 80% Cultivation					
1659	8930691	2060	11109848	24.17%	24.40%
STRATUM 12 – 51% - 80% Cultivation					
2046	10941340	1870	9853959	- 8.60%	-9.94%
STRATUM 20 – 15% - 50% Cultivation					
1816	9100479	1746	8585013	-3.85	-5.66%
STRATUM 40 – Less than 15% Cultivation					
2104	20534800	1949	19958490	-7.37	-2.81%

4.2. Corn, Soybean and Winter Wheat Estimates

To test this hypothesis, state level corn, soybean and winter wheat estimates for 270 samples were obtained using the Nebraska area frames stratified using 2015 CDLs created with 30 m or non-buffered training data. The resulting crop specific estimates are plotted in Fig. 3 which includes histograms that illustrate the sample distribution of the crop estimates obtained for Nebraska 2015 corn, soybean and winter wheat.

The crop estimate results, illustrated in Fig. 3 and Table 3 indicate that using a 30 m buffer vs. non buffer for area frame stratification and estimation results in crop estimates that are statistically significantly different. The corn, soybean and winter wheat estimates were consistently lower based on the CDLs created using the 30 m buffered training data (Fig. 3 – bottom row), when compared to the crop estimates obtained using CDLs created with non-buffered training data (Fig. 3 – top row). The results of a t-test conducted using the Pooled and Satterthwaite methods to test the difference in the sample distribution means for each of the crop estimates are included in Table 3. At the 95% confidence interval, there was a statistically significant difference in the mean for all crops and for the two buffer types with all p-values less than .0001. Table 4 includes the test for differences in variances conducted using the Folded F method. There was no significant difference in variance between all pairs.

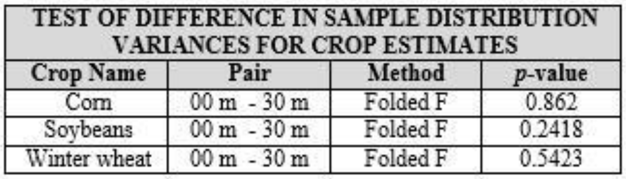
TABLE 3. Test of Difference in Crop Estimate Means



TEST OF DIFFERENCE IN SAMPLE DISTRIBUTION MEANS FOR CROP ESTIMATES

Crop Name	Pair	Method	*p*-value
Corn	00 m - 30 m	Pooled	<.0001
Corn	00 m - 30 m	Satterthwaite	<.0001
Soybeans	00 m - 30 m	Pooled	<.0001
Soybeans	00 m - 30 m	Satterthwaite	<.0001
Winter wheat	00 m - 30 m	Pooled	<.0001
Winter wheat	00 m - 30 m	Satterthwaite	<.0001

TABLE 4. Test of Difference in Crop Estimate Variances



TEST OF DIFFERENCE IN SAMPLE DISTRIBUTION VARIANCES FOR CROP ESTIMATES

Crop Name	Pair	Method	*p*-value
Corn	00 m - 30 m	Folded F	0.862
Soybeans	00 m - 30 m	Folded F	0.2418
Winter wheat	00 m - 30 m	Folded F	0.5423

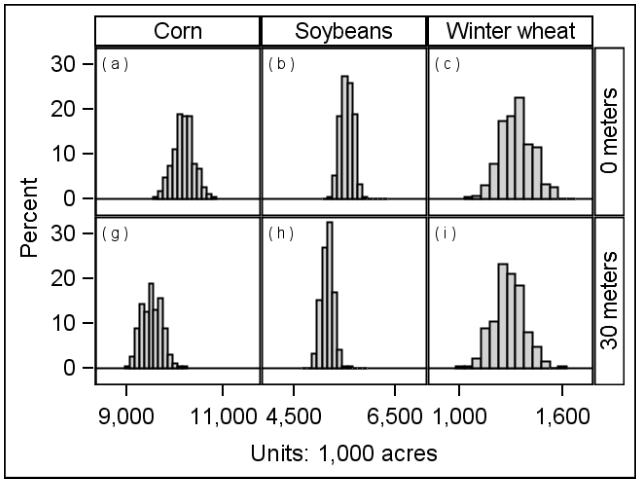


Fig. 3. Histogram of Nebraska 2015 corn, soybean and wheat estimates. Crop estimates are obtained using CDLs created with 30 m buffer or no buffer training data for area frame stratification and crop estimation.

5. CONCLUSION

This study evaluates the impact of training data pixel level buffering on Nebraska area frame stratification results and crop estimates. Results show that pixel level buffering of FSA CLU training data has a large impact on area frame stratification results; and that using 30 m buffered training data for CDLs has a statistically significant impact on Nebraska 2015 corn, soybean and winter wheat estimates. Results show that using the 30 m buffered CDL for area frame stratification and crop estimation yields statistically lower corn, soybean and winter wheat estimates. These results may explain why current CDLs, produced using a 30m buffer, consistently underestimate crop acreage [8].

6. REFERENCES

[1] Cotter, J. C. Davies, J. Nealon, and R. Roberts.,*Area Frame Design for Agricultural Surveys in Agricultural Survey Methods* (eds R. Benedetti, M. Bee, G. Espa and F. Piersimoni), John Wiley & Sons, TD, Chichester, UK, 2010.

[2] Ford, B. I., J. Nealon, R.D. Tortora, R, “Area frame estimators in agricultural surveys; sampling versus non sampling errors,” *Agricultural Economics research* 38 (2), 1-10. 1986.

[3] Vogel, F. A, “The evolution and development of agricultural statistics at the United States Department of Agriculture, “Journal *of Official Statistics*, 11: 161-180. 1995.

[4] Nusser S and C. House, *Sampling, data collection, and estimation in agricultural surveys.* In: Pfeffermann D, Rao C (eds) Handbook of statistics 29A, sample surveys: design, methods and applications. Elsevier. The Netherlands, pp 471-486. 2009

[5]. Boryan, C. Z. Yang, R. Mueller, and M. Craig, “Monitoring US Agriculture: The US Department of Agriculture, National Agricultural Statistics Service Cropland Data Layer Program,” *Geocarto* International 26, (5): 341-358, 2011.

[6] Boryan, C., Z. Yang, L. Di., and K. Hunt, “A New Automatic Stratification Method for U.S. Agricultural Area Sampling Frame Construction Based on the Cropland Data Layer,” IEEE *Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, pp. 1939-1404, Nov. 2014, DOI: 10.1109/JSTARS.2014.2322584.

[7] Boryan, C. and Z. Yang, “Operational implementation of a new automatic stratification method using geospatial cropland data layers in the NASS area frame section”, , proceedings of IGARSS 2014, IGARSS 2014 & 35th Canadian Symposium on Remote Sensing, Quebec City, Canada, July 13-18, 2014.

[8] Lark, Tyler, “Measuring Land - Use and Land - Cover Change using the USDA Cropland Data Layer: Cautions and Recommendation”, Presented at American Geophysical Union, Dec 12 – 16, San Francisco, California, 2016

[9] FSA, Farm Service Agency Common Land Unit Information worksheet.<http://www.fsa.usda.gov/Internet/FSA_File/clu__infosheet_2013.pdf> (last accessed 20, Dec 2016), 2016.

[10] Yang, Z., C. G. Boryan, M. Hyman, P. Willis, “The Impact of Single-Pixel Ground Truth Field Boundary Buffering on Medium Resolution Land Cover Classification Accuracy”, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing Special Agro-Geoinformatics Issue*. Submitted, 2016.

[11] Boryan, C., Z. Yang, P. Willis, and R. Hardin, "Evaluating the impact of pixel level buffering on area sampling frame stratification results”, Presented at Fifth International Conference on Agro-geoinformatics, July 18 – 20, Tianjin, China, 201

[12] NASS. “Nebraska 2015 Cropland Data Layer Metadata https://www.nass.usda.gov/Research_and_Science/Cropland/metadata/metadata_ne15.htm (last accessed 21, Dec 2016), 2016.