A new land cover classification based stratification method for area sampling frame construction

Claire G. Boryan, Zhengwei Yang
USDA National Agricultural Statistics Service
3521 Old Lee Highway, Room 305, Fairfax, VA 22030, U.S.A
Email: Claire.Boryan@nass.usda.gov

Abstract—**This paper proposes a new automated USDA National Agricultural Statistics Service (NASS) Cropland Data Layer (CDL) based method for stratifying U.S. land cover. The proposed method is used to stratify the NASS state level Area Sampling Frames (ASFs) by automatically calculating percent cultivation at the Primary Sampling Unit (PSU) level based on the CDL data. The CDL based stratification experiment was successfully conducted for Oklahoma, Ohio, Virginia, Georgia, and Arizona. The stratification accuracies of the traditional and new automated CDL stratification methods were compared based on 2010 June Area Survey (JAS) data. Experimental results indicated that the CDL based stratification method achieved higher accuracies in the intensively cropped areas while the traditional method achieved higher accuracies in low or non agricultural areas. The differences in the accuracies were statistically significant at a 95% confidence level. It is concluded that the CDL based stratification method will improve efficiency and reduce cost in NASS ASF construction, and improve the precision of NASS JAS estimates.**

Keywords-stratification; land cover; area sampling frame; primary sampling unit; CDL

I. INTRODUCTION

Area Sampling Frames (ASFs) are the foundation of the agricultural statistics program of the National Agricultural Statistics Service (NASS) and many other statistical survey programs. ASFs have been used since 1954 as a primary tool for conducting surveys to gather information on crop acreage and other agricultural information. They are considered "the backbone to the agricultural statistics program of the National Agricultural Statistics Service (NASS)" [3]. NASS' primary area frame survey is the June Area Survey (JAS) in which 11,000, one square mile sample segments, are visited by enumerators each year at the beginning of the growing season to collect crop type and acreage information. Estimates of crop acreage and livestock inventories are based on the data collected during the JAS. The NASS ASFs are based on the stratification of U.S. land cover into homogeneous groups or strata based on percent cultivation. This stratification of land cover has been conducted using visual interpretation of aerial or satellite data for the past fifty eight years. The precision and accuracy of the survey statistics are dependent on the techniques used to construct and sample the NASS Area Frame.

NASS, also, has a remote sensing acreage estimation program in which satellite data acquired throughout the growing season are utilized as inputs to produce crop specific land cover classifications from which independent acreage estimates are generated. These state level agricultural land cover classifications are known as Cropland Data Layer (CDL) products. The NASS CDLs are 30-56.0 meter raster-formatted, geo-referenced, crop-specific land cover classifications. Historically, CDLs were produced, beginning in 1997, for major crop producing states in the Midwest of the United States to provide acreage estimates to the NASS Agricultural Statistics Board (ASB) and Field Offices (FOs). Over the years, the program has expanded to include all 48 US conterminous states for years 2008-2011. Total crop mapping accuracies for historic CDLs ranged from 85% to 95% for the major crop categories. Boryan et al., [1] provide greater detail regarding CDL production. The CDL data are publically available from NASS' online geospatial application - CropScape [2].

Currently, NASS' Research and Development Division is working to improve the efficiency, reduce the cost and improve the precision of the estimates generated from the June Area Survey. Toward this goal, a new automated method has been developed to objectively, consistently, and rapidly stratify U.S. land cover, based on percent cultivation, of the Area Sampling Frame (ASF) Primary Sampling Units (PSUs) using the 2010 CDL products.

This paper proposes a new automated, NASS CDL based stratification method and makes a performance comparison between the NASS traditional method which is based on visual interpretation and the new automated CDL stratification method. The effectiveness of the traditional and CDL based methods in determining percent cultivation, at the Area Frame Primary Sampling Unit (PSU) level, were assessed using in situ validation data collected at the segment level as part of the 2010 JAS. The goal of this investigation was to determine the utility of the automated CDL based method for use in the stratification of U.S. land cover and potentially in ASF construction [3]

II. BACKGROUND

A. NASS Area Sampling Frames

The NASS ASFs are based on a stratification of land cover in the U.S. by percent cultivated cropland, and are the statistical foundation for providing estimates with complete coverage of U.S. agriculture. The Area Frame program is

conducted in 49 states using approximately 11,000 one square mile segments made up of approximately 41,000 individual farms. Selected farms are visited each year by enumerators, as part of the JAS, to identify the planting intentions for all agricultural land within the segments, including planted acreage and acreage intended for harvest. Acreage estimates for major commodities such as corn, soybeans, winter wheat, spring wheat, durum wheat and cotton are generated from the JAS at the state and national level.

The primary use of the ASF within NASS is as the foundation of the JAS. The ASF is also used to measure the incompleteness of NASS list frame, provide ground truth for the NASS Cropland Data Layer (CDL) program to generate independent acreage estimates, and for additional surveys such as NASS' objective yield survey and the Agricultural Coverage Evaluation Survey.

Area frame construction is a lengthy process conducted one state at a time. Frames are generally in use for approximately 15 to 20 years with some in operation for as long as 30 or more years. Fig. 1 illustrates the implementation years of current NASS state level Area Frames. Originally when ASFs were created on paper, only two frames were built per year. Currently three to four frames are built each year due to technological improvements including the use of ESRI's ArcGIS software, aerial photography, satellite imagery and ancillary agricultural information when available. On average, five full time employees working for a period of four months per state are required for new frame construction [3].

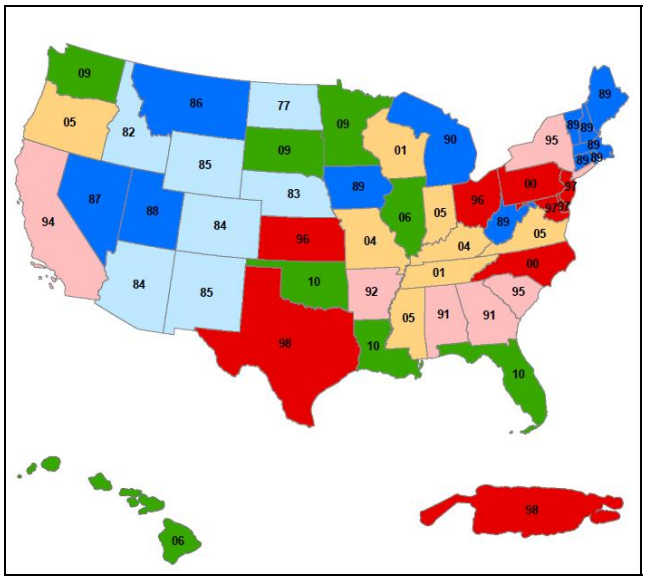

Figure 1. US map illustrating the implementation years of current NASS Area Frames

Area Sampling Frames (ASFs) have been used in a variety of research applications including an evaluation of the prevalence of brown stem rot in the north central United States [4], to improve agricultural ground survey estimates as part of the Monitoring Agriculture with Remote Sensing (MARS) project launched by the European Union in 1989 [5] and to developed a Geographic Information System for crop area estimation at a regional level in the Islamic Republic of Iran [6]. Faulkenberry and Garoui [7] compared the utility of frequently used estimators based on an ASF utilized in agricultural surveys.

B. NASS Area Sampling Frame: Land Use Stratification

Land use stratification is "the division of land area into broad land use categories" and is known to improve efficiency for statistical sampling and estimation. In the construction of a NASS area sampling frame (ASF), general cropland (based on percentage cultivation), agriculture/urban, residential/ commercial, and non agriculture are the commonly identified strata. The agricultural strata definitions vary between states depending on the type and intensity of agricultural production. Strata homogeneity is critical for the performance of the NASS ASFs. Once strata definitions are assigned, all land is subdivided into primary sampling units (PSUs). Specific PSUs are allocated for inclusion in JAS. These PSUs are further subdivided into segments, and a segment is randomly selected from each allocated PSU for enumeration [3].

TABLE I. Land-use stratification codes and definitions represented in the NASS Area Sampling Frames (Benedetti, 2010)

Land-Use Strata Codes (Stratum)	Strata Definition
11	General Cropland, greater than 75% cultivated.
12	General Cropland, 51-75% cultivated
20	General Cropland, 15-50% cultivated.
31	Ag-Urban, less than 15% cultivated, more than 100 dwellings per square mile, residential mixed with agriculture.
32	Residential/Commercial, no cultivation, more than 100 dwellings per square mile.
40	Less than 15% cultivated
50	Non-agricultural,
62	Water

For the past 58 years, land use stratification has been conducted using visual interpretation of satellite imagery or aerial photography. Most recently, the satellite data used was acquired by the Landsat Thematic Mapper (TM), which was primarily relied upon to identify cultivated cropland and, when necessary, identify specific crop types. The National Agricultural Imagery Program (NAIP) data; which are one meter, ortho-rectified, air photos acquired during the growing season; are utilized as the base for digitizing PSU boundaries.

C. The NASS Cropland Data Layer Products

The NASS Cropland Data Layer (CDL) products are 30-56.0 meter, raster-formatted, geo-referenced, crop-specific land cover classifications. The first state level CDL was produced in 1997 and CDLs for all 48 conterminous states in the U.S were produced from 2008- 2011. The purpose of the CDL program is to use satellite data to provide acreage estimates for important crop producing states to the NASS Agricultural Statistics Board (ASB). All historic CDL products are publically available on-line for accessing, visualization, downloading, map printing, as well as on the fly statistical and change analysis from the NASS' web service based CDL application – CropScape [2].

The 2010 CDLs were produced for all 48 states at a 30 meter spatial resolution as illustrated in Fig. 2. The 2010 CDLs were produced using satellite data acquired, during the growing season, from the Indian RESOURCESAT-1 (IRS-P6) Advanced Wide Field Sensor (AWiFS), Landsat 5 Thematic Mapper (TM) and Landsat 7 Enhanced Thematic Mapper Plus (ETM+). Ancillary data inputs were also used. They included: the United States Geological Survey (USGS) National Elevation Dataset (NED), the USGS Percent Canopy layer, the USGS Percent Impervious layer, and the National Aeronautics and Space Administration (NASA) Moderate Resolution Imaging Spectroradiometer (MODIS) 250 meter 16 day Normalized Difference Vegetation Index (NDVI) composites. Farm Service Agency (FSA) Common Land Unit (CLU) data were used for training and validation of agricultural categories. The USGS National Land Cover Dataset (NLCD) 2001 was used as training and validation for the non agricultural categories. In general, total crop mapping accuracies for the 2010 CDLs ranged from 85% to 95% for the major crop categories [1]. The accuracies for major crops, such as corn and soybean in major US agricultural areas are in the high 90th percentile. These accuracy numbers are considered high for stratification purposes as errors between crop categories can be canceled in the stratification process.

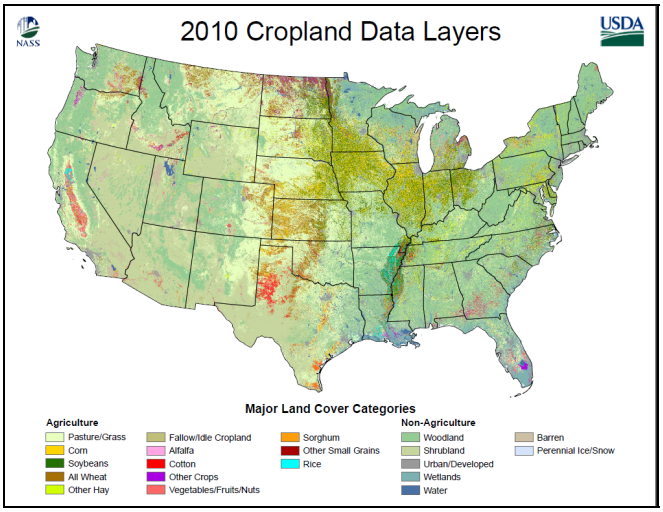

Figure 2. 2010 NASS Cropland Data Layer products.

III. METHODOLOGY

A. *Data and Preprocessing*

The data files required for each of the five state analyses conducted in this investigation included: a state level ASF with strata specific PSUs, a 2010 state level NASS CDL, and a 2010 JAS segment file with segment level percent cultivation calculated. The 2010 CDLs included as many as fifty categories and a wide variety of crops. For the purpose of stratification, a crop mask [8] was first generated by recoding CDL pixels of crop categories into "cultivated" and non crop pixels into "non-cultivated". To build the 2010 CDL crop masks, crop and non-crop categories were recoded to "1" and "0" respectively. A typical crop mask is illustrated in Fig.3 (a) with green and white colors representing the cultivated and non-cultivated pixels respectively. The crop mask is overlaid with an ASF with PSU boundaries. The exact crop types in each PSU are illustrated in Fig. 3(b) by overlaying the ASF over the CDL. The results of the traditional stratification and the corresponding CDL based stratifications were compared based on the JAS segment in situ ground truth.

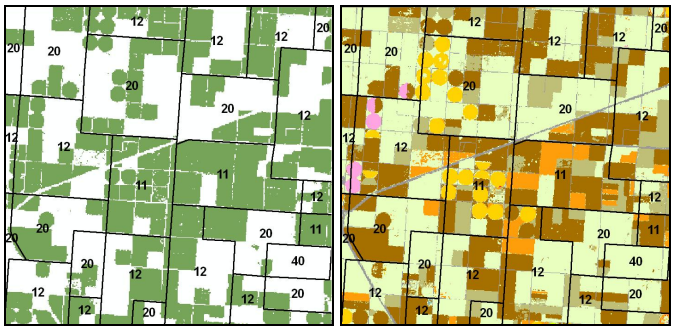

Figure 3. (a) An ASF with PSU boundaries overlaying a CDL crop mask (green- cultivated, white- non cultivated); (b) the same ASF overlaying a 2010 CDL image product (brown – winter wheat, yellow, corn, orange – sorghum, pink- alfalfa, pale green – other hay, blue-water).

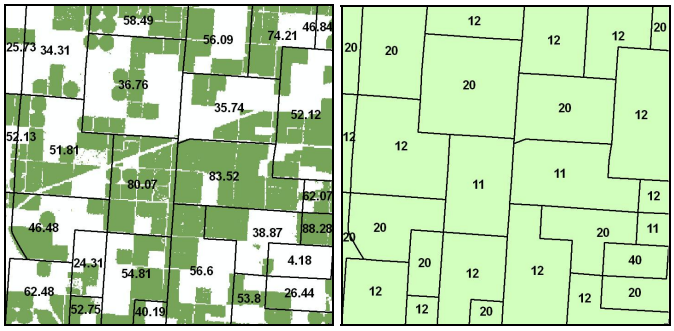

Figure 4. (a) CDL crop mask based PSU percent cultivation; (b) an ASF with CDL derived stratification.

B. *Stratification Method*

NASS' traditional area frame stratification process involves visual interpretation of ASF PSUs into different strata based on percent cultivated land within a PSU boundary. The percent cultivation of each ASF PSU can be calculated from state level crop masks derived from CDL by counting pixels with value "1" (cultivated) and total number of pixels within the PSU boundary. The percent cultivated is given by the number of "1" pixels divided by total number of pixels. Fig. 4(a) illustrates PSU percent cultivation calculated from CDL based crop mask. With the calculated percent cultivation, each PSU's stratum could be labeled into a different stratum category based on the state specific strata definitions as given by Table 1. The resulting CDL based stratification for the same frame is shown as Fig. 4(b). This proposed new CDL based process is objective and automated while the traditional stratum labeling method is subjective and manual. This new method improves efficiency, objectiveness and accuracy in stratification.

TABLE II. Oklahoma 2010 ASF Analysis, Traditional vs. CDL Stratification Method

		Traditional Stratification			CDL Stratification			p-value
Stratum	*% Cultivated*	*Segments*	*Correct*	*Accuracy (p_1)*	*Segments*	*Correct*	*Accuracy(p_2)*	*Ha: $p_1 \neq p_2$*
11	>75%	140	47	34%	43	27	63%	0.001
12	51% - 75%	48	9	19%	77	30	39%	0.024
20	15% - 50%	74	26	35%	98	42	43%	0.305
40	< 15%	61	61	100%	105	96	91%	0.027
Total		323			323			

TABLE III. Ohio 2010 ASF Analysis, Traditional vs. CDL Stratification Method

		Traditional Stratification			CDL Stratification			p-value
Stratum	*% Cultivated*	*Segments*	*Correct*	*Accuracy (p_1)*	*Segments*	*Correct*	*Accuracy(p_2)*	*Ha: $p_1 \neq p_2$*
11	>75%	110	84	76%	85	76	89%	0.019
12	51% - 75%	35	15	43%	42	23	55%	0.055
20	15% - 50%	42	28	67%	48	37	77%	0.271
40	< 15%	53	47	89%	65	57	88%	0.869
Total		240			240			

TABLE IV. Five State 2010 Strata Summary, Traditional vs. CDL Stratification Method

		Traditional Stratification			CDL Stratification			p-value
Stratum	*% Cultivated*	*Segments*	*Correct*	*Accuracy (p_1)*	*Segments*	*Correct*	*Accuracy(p_2)*	*Ha: $p_1 \neq p_2$*
11	>75%	250	131	52%	128	103	80%	0.000
12	51% - 75%	83	24	29%	119	53	45%	0.025
13	>50%	171	90	53%	91	69	76%	0.000
20	15% - 50%	371	177	48%	387	219	57%	0.000
40	< 15%	322	305	95%	472	407	86%	0.000
Total		1197	727	61%	1197	851	71%	0.000

C. Stratification Result Evaluation

To evaluate the effectiveness of the CDL based stratification method, the results were compared with the traditional stratification method based on 2010 JAS segment ground truth data. Stratification performance was assessed based on the percent of segments matching the definition (correctly labeled stratum) among all segments labeled with a given stratum of the PSUs within which they are located. The reference segment data were derived from JAS survey. The enumerators recorded ground truth in the survey.

An assessment of the resulting accuracies indicated that the CDL stratification method generally resulted in higher accuracies, but these results did not indicate whether the differences in the two methods were significant statistically. The ultimate goal of this evaluation was to determine whether the proposed CDL based stratification method achieved equivalent or improved accuracies when compared to the traditional Area Frame stratification method. Therefore, two-tailed proportion tests, a Chi-Square test or a Fisher's Exact test for sample sizes less than five were conducted for each state and each stratum. In these tests, two sample proportions were p_1 the accuracy results from the traditional stratification method and p_2 the accuracy results from the proposed CDL based stratification method. The hypotheses of the significance tests were H_0: $p_1=p_2$ and H_a: $p_1 \neq p_2$. The null hypothesis stated that there was no difference in the accuracies of the two stratification methods while the alternative hypothesis stated that the accuracies of the two stratification methods were significantly different. The tests were performed and p values were calculated for each state and each stratum with a confidence level of 95%.

IV. Discussion

A comparison was made between the stratification results achieved by the traditional stratification method and those of the CDL based method using JAS reported data as in situ validation. It should be noted that the CDL stratification method could only identify percent cultivation, non agriculture, and water. ASFs often have additional strata such as stratum 31 (Ag-Urban, less than 15% cultivated, more than 100 dwellings per square mile) and stratum 32 (residential/ commercial, no cultivation, more than 100 dwelling per square mile). For this analysis, these strata were included in stratum 40 (less than 15% cultivated). All stratification was PSU based.

Table 2 and Fig. 5 presented the accuracy results for both the traditional stratification of the ASF PSUs and those of the CDL derived stratification of the ASF PSUs for 2010 Oklahoma frame analysis. All other states were evaluated in the same manner. In Tables 2, 3, and 4, the accuracy results of the traditional (visual interpretation) stratification method and the CDL based stratification method were presented side by side. The column at the far right identifies the p-values of the Chi Square or Fisher's Exact tests.

As shown in Table 2, for the Oklahoma 2010 traditional stratification, of the 323 total segments in the state, 140 were in PSUs that were identified as being stratum 11 (greater than 75% cultivated) by NASS carto-technicians based on stratum definition. Of the 140 segments, JAS reported that 47 were stratum 11, an accuracy of 34%. For the CDL stratification results, of the 323 segments, 43 were labeled in stratum 11 PSUs. Of the 43 segments, the JAS reported 27 segments were

stratum 11, an accuracy of 63%. The accuracies of the CDL stratification method were higher than those of the traditional area frame stratification method.

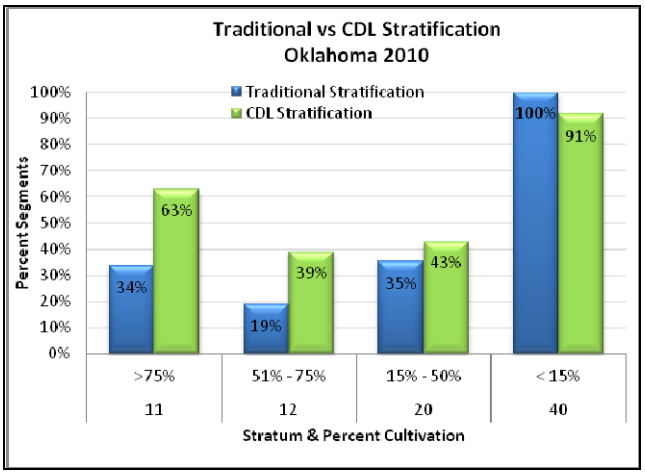

Figure 5. Oklahoma 2010 ASF Analysis, Traditional Method vs. CDL Based Stratification Method

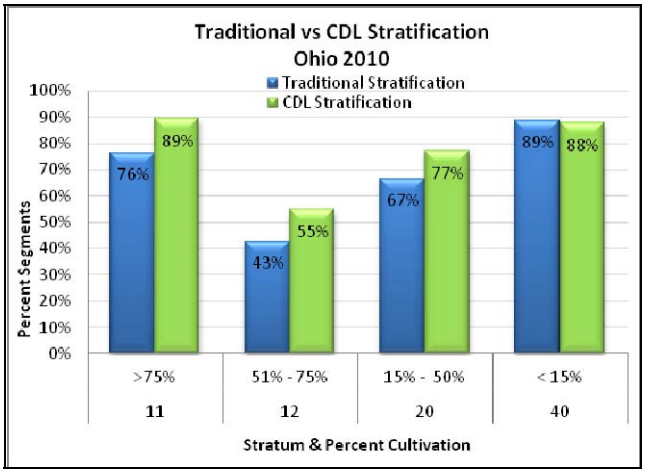

Figure 6. Ohio 2010 ASF Analysis, Traditional Method vs. CDL Based Stratification Method

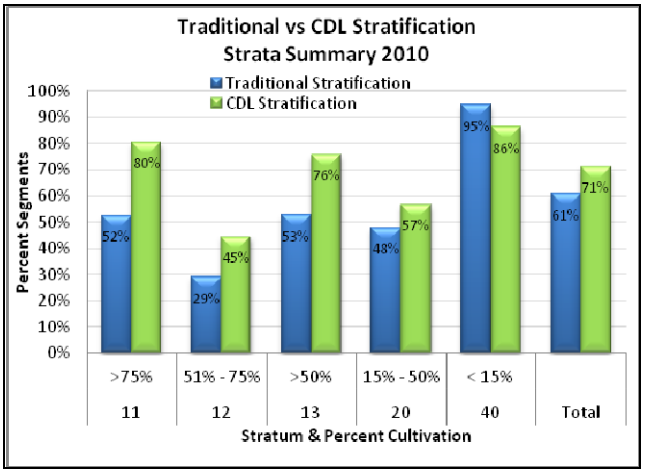

Figure 7. Five State 2010 Strata Summary, Traditional Method vs. CDL Based Stratification Method

Similarly, in Stratum 20 (15% - 50% cultivated), 74 segments were in PSUs identified by the traditional stratification as stratum 20. Of these 74 segments, the JAS reported that 26 were stratum 20, which represented an accuracy of 35%. Of the 323 segments, 98 were located in PSUs identified by the CDL based stratification as stratum 20. Of the 98 stratum 20 segments, the JAS reported that 42 were stratum 20, an accuracy of 43%. The accuracies of the CDL stratification method were higher than those of the traditional stratification method.

As shown in Table 2, the accuracies of the traditional stratification, the CDL based stratification and the JAS survey reported were calculated for the Oklahoma ASF and summarized in the same manner for all strata. In Table 2, the p-values were highlighted in red if the differences were considered statistically significant at the 95% confidence level. Out of the eight comparisons, six were significantly different from one another. Stratum 20 was the only stratum that showed no statistically significant difference between the two stratification methods at the confidence level of 95% although the accuracies of the CDL stratification method were higher than those of the traditional stratification method.

Note that in Table 2 the accuracies are derived for Oklahoma segments meeting formal stratum definitions using the traditional and CDL based stratification methods. Tables 2 and 3 illustrate the results for the Oklahoma and Ohio analyses. Figs. 5 and 6 provide graphical representations of the Oklahoma and Ohio results.

Table 4 and Fig. 7 provide a summary of all state level analyses across five strata. In summary, land cover in the five states stratified with the automated CDL method has a higher rate of accuracy than that of the traditional method when using the JAS survey data as validation. Across all ASF strata specifically, 11, 12, 13, 20 and 40, the traditional method achieved a total accuracy of 61% and the CDL stratification method achieved a total accuracy of 71%, as shown in Table 4. The CDL stratification method achieved higher accuracies in all strata except stratum 40 (Traditional Method - 95% vs. CDL method - 86%). The CDL method achieved higher accuracy in the more highly cultivated strata such as stratum 11 (Traditional Method - 52% vs. CDL Method - 80%) and stratum 13 (Traditional Method - 53% vs. CDL Method -76%).

Two sided proportion tests were conducted on all comparisons to test whether the differences in accuracies achieved using the two different methods were statistically significant. In these tests the significance level was set to be 0.05 and the null hypothesis was accepted if the *p*-value exceeded the significance level. Across all strata, when state results were combined, the *p*-values were less than 0.05. The tests indicated that in strata 11, 12, 13 and 20 the CDL stratification method was more accurate at determining percent cultivation. In stratum 40, the traditional method was more accurate.

As observed in Tables 2, 3 and 4, the CDL method achieved higher rates of accuracy in the highly cultivated strata (11, and 13). Identification of cropland is the strength of the CDL process. Satellite imagery collected across the entire growing season, decision tree classification software, useful

ancillary data and an abundance of training data enable the CDLs to achieve total crop accuracies of approximately 85 – 95%. Furthermore, the CDL method is objective and consistent.

Both methods struggled in having the identification of percent cultivation for the PSUs match the JAS segment level data in strata 12 and 20. The primary issue with these strata is due to the heterogeneity of the land cover. In the construction of the ASF, cartographic technicians attempt to define PSUs that are homogeneous. It is very difficult to define a stratum 12 or 20 PSU that is homogeneous across the PSU. The cropland generally is clustered in one portion of the PSU and the remainder of the PSU has small amounts of agriculture. When the segments are selected in these PSUs it is not unusual that they not represent their strata definition. One recommendation to improve PSU homogeneity would be to reduce the size of PSUs during new ASF construction. This would improve ASF performance in strata 12 and 20 PSUs. One area of further research for this project is to determine if the CDL based method can more accurately identify percent cultivation in the strata 12 and 20 PSU using multi-year data rather than single year CDL data.

V. CONCLUSION

This paper proposed a new automated CDL based method for deriving percent cultivation and subsequently stratifying U.S. land cover. The CDL based stratification of NASS ASF PSUs was successfully conducted for Oklahoma, Ohio, Virginia, Georgia, and Arizona. The stratification accuracies of the traditional and new CDL based stratification methods were compared based on in situ validation data collected by enumerators during the 2010 JAS. Results of the five state analyses indicated that the new automated CDL method was more accurate in determining U.S. percent cultivation in intensively cropped areas and weaker in non agricultural areas. The CDL based stratification achieved higher accuracies in strata 11, 12, 13 and 20 while the traditional method achieved higher accuracies in stratum 40. The differences in the accuracies were statistically significant at a 95% confidence level. The novelty of the proposed method is using geospatial CDL data to objectively and automatically compute percent cultivation of the ASF PSUs as compared to the traditional method that subjectively determining percent cultivation based on visual estimation of satellite data. This proposed new CDL based process improved efficiency, objectiveness and accuracy as compared to the traditional stratification method. It is concluded that adoption of the automated CDL stratification method in ASF construction will help NASS achieve the goals of improving efficiency, reducing cost and improving the precision of JAS estimates by updating the NASS ASFs with greater frequency and stratification accuracy.

REFERENCES

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

[2] Han, W., Yang, Z., Di, L., Mueller, R., "CropScape: A Web service based application for exploring and disseminating US conterminous geospatial cropland data products for decision support." Computer and Electronics in Agriculture, Vol. 84, June, pp. 111–123, http://dx.doi.org/10.1016/j.compag.2012.03.005

[3] Benedetti, R., Bee, M., Espa, G., Piersimoni, F. Cotter, J. Davies, C., Nealon, J. Roberts, R., 2010. Agricultural Survey Methods; Chapter 11. Area Frame Design for Agricultural Surveys. John Wiley & Sons, Ltd. Published Online: 25 March 2010.

[4] Workneh, F., Tylka G., Yang, X., Faghihi, J. and Ferris J., "Regional Assessment of Soybean Brown Stem Rot, Phytophthora sojae, and Heterodera glycines Using Area-Frame Sampling: Prevalence and Effects of Tillage," Phytopathological, March 1999, Volume 89, Number 3, Pages 204-211.

[5] Tsiligirides, T., "Remote sensing as a tool for agricultural statistics: a case study of area frame sampling methodology in Hellas," Computers and Electronics in Agriculture, Vol. 20 (1998) pp. 45–77.

[6] Pradhan, S. 2001, "Crop area estimation using GIS, remote sensing and area frame sampling," International Journal of Applied Earth Observation and Geoinformation, Volume 3, Issue 1, 2001, Pages 86-92.

[7] Faulkenberry, G. D. and Garoui, A. 1991, "Estimating a Population Total Using an Area Frame," Journal of the American Statistical Association, Vol. 86, No. 414 (Jun., 1991), pp. 445-449.

[8] Boryan, C., Yang, Z., and Di. L., "Deriving 2011 cultivated land cover data sets using usda national agricultural statistics service historic cropland data layers," Proc. of IEEE International Geoscience and Remote Sensing Symposium, July 22-27, 2012, Munich, Germany.