DERIVING CROP SPECIFIC COVARIATE DATA SETS FROM MULTI-YEAR NASS GEOSPATIAL CROPLAND DATA LAYERS

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

The National Agricultural Statistics Service (NASS) Area Sampling Frames (ASFs) are based on the stratification of US land cover by percent cultivation. Recently, an automated stratification method based on the NASS Cropland Data Layer (CDL) was developed to efficiently and objectively stratify US land cover. This method achieved higher accuracies in all cultivated strata with statistical significance at a 95% confidence level. This paper proposed to develop crop specific covariate data based on 2007 – 2010 CDLs. Crop (corn, soybeans, wheat and cotton) and non crop (forest, urban and water) covariate data were derived and validated for six states. Producer and user accuracies for the covariate data sets were based on independent 2011 Farm Service Agency Common Land Unit data and 2011 CDLs. Non crop covariate data were validated using the National Land Cover Data 2006. Covariate data were used within NASS to conduct substratification of the 2013 Oklahoma ASF.

Index Terms— *Cropland Data Layer, crop covariate data, stratification, area sampling frame*

1. INTRODUCTION

The USDA National Agricultural Statistics Service (NASS) has produced state level Cropland Data Layers (CDLs) for major US agricultural states since 1970 and for all 48 conterminous states since 2009 [2]. NASS uses CDLs to directly produce acreage estimates of major crops in agriculturally intensive states. June Agricultural Survey (JAS) segments are used to perform a simple linear regression to derive the crop specific acreage estimates [2]. Recently, Boryan and Yang used CDL data in Area Sampling Frame (ASF) construction to automatically stratifying land cover in the US based on percent cultivation [4]. It was found that the automated CDL based stratification significantly improved stratification accuracies in intensively cropped areas and performed less well in non agricultural areas as compared with the visual interpretation based traditional stratification method for five test states. The differences in accuracies were statistically significant at a 95% confidence level [4].

To further improve crop estimates, including additional covariates may be predictive of where crops will be grown in the future. The ASFs are built for future use. Therefore, to improve area sampling design and ultimately survey estimation for individual major crops, the ASF Primary Sampling Units (PSUs) have to substratified based on crop specific information rather than solely on percent cultivation, i.e. the crop specific covariates have to be derived. In the past, the ASF PSUs were substratified based on county estimates for specific crops [1]. The covariate data sets provide the unique opportunity to derive percentages for these crops at the PSU level. This paper describes a new procedure for deriving crop specific covariate data sets (the cultivated crop land of specific crops types) and non crop categories of interest from 2007 – 2010 multi-year CDL data. The capability of the derived crop specific covariate data to predict the amount of cultivated crop land and specific crops types in a future year, 2011, was assessed.

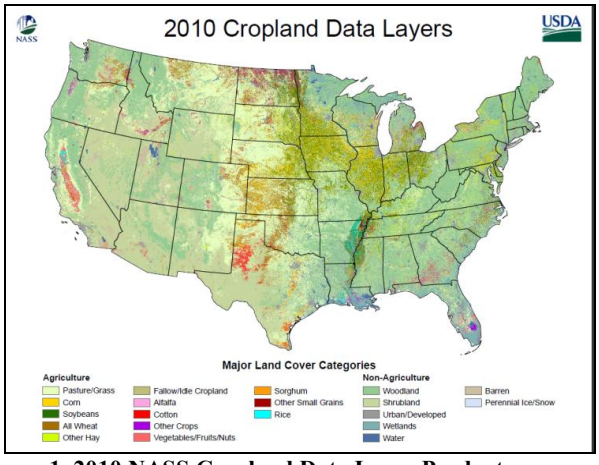

Figure 1. 2010 NASS Cropland Data Layer Products

2. DATA AND METHODOLOGY

The covariates of interest including corn, soybeans, wheat, cotton, forest, water and urban were derived from multiyear (2007-2010) CDL data. They were used to predict their presence in 2011. The 2007-2010 final CDLs for California, Indiana, Mississippi Nebraska, Pennsylvania and Washington were used as the inputs to develop the covariate data sets for this assessment. To develop the covariate data, all CDL data were resampled to 30 meters using nearest neighbor interpolation. All variables were recoded into a set of new codes different from the original CDLs. In composing all variables of interest, land cover categories were merged. For example, when building the forest variable, all pixels in the original CDLs that were identified to deciduous, coniferous or mixed forest were merged together.

2.1 Rules for deriving covariates

A set of rules was developed to derive the data from the multi-year CDL data. The set of rules defining all major variables are given in the following subsections.

2.1.1. Cultivated variable

All of the original CDLs and validation data were recoded from their original categories to cultivated and non cultivated. In NASS, cultivated land is specifically identified as land that is used for growing crops and includes fallow or idle land. It does not include grass, pasture or non alfalfa hay. To composite the cultivated variable from the multi-year CDLs, two model rules were defined: (1) all pixels ever categorized to a cultivated crop in all available years are retained; (2) pixels categorized to a cultivated crop at least two times in the original CDL inputs are retained [3]. The appropriate models were chosen for each test state based on average state CDL accuracy [7]. CDL accuracy information is available on the NASS Research Division web site [7]. As shown in Fig. 2, NASS, based on previous research [3], produced a cultivated layer from multi-year (2008-2012) CDL data using these defined models. The cultivated layer is available to the public at:
http://www.nass.usda.gov/research/Cropland/Release.
Accuracies for the cultivated category are provided at the same site.

2.1.2. Crop (corn/soy, cotton, wheat) variables

All pixels categorized to these crops in all available years are retained in the crop specific data sets. The corn/soybean category is the only category in which two crops types are included in the same category. In many states, corn and soybeans are commonly rotated so they are considered one category for this assessment.

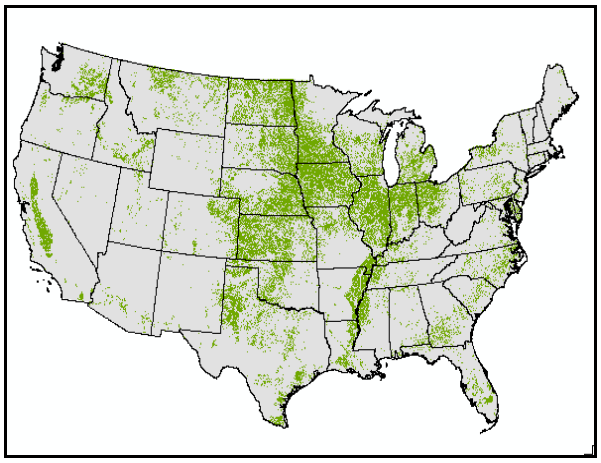

Figure 2. 2012 NASS Cultivated Layer

2.1.1. Non Crop (forest, water, and urban) variables

All pixels categorized to the specific category more than two times in the original CDL inputs are retained.

2.2 Automated covariate stratification method

The crop covariate data sets are used as the input to an automated stratification procedure [3]. The GIS procedure automatically and objectively determines the percent crop at the ASF PSU level using the multi-year CDL land cover information. The detailed steps are given as follows:

1) Derive state level covariate data sets from multi-year (2007-2010) CDL data by combining the specific crop (i.e. corn/soy, wheat or cotton) over the multi-year period into one crop category and assigning the corresponding pixels with a value of "1" while grouping the rest of categories into one "other" category and assigning the corresponding pixels with a value of "0";
2) Load an individual ASF PSU boundary;
3) Load a CDL covariate layer
4) Overlay an ASF PSU boundary on the CDL covariate layer;
5) Compute percent covariate of each ASF PSU by counting the total number of pixels with value "1" (specific crop) and the total number of all pixels within the PSU boundary. The percent covariate is given by the number of "1" pixels divided by total number of pixels.

3. RESULTS AND DISCUSSION

The cultivation data sets were validated using 2011 Farm Service Agency (FSA) Common Land Unit (CLU) data and 2006 National Land Cover Data(NLCD). The corn/soybeans cotton and wheat data sets were validated based on the 2011 CDLs. The forest, water and urban data sets were validated based on the NLCD, 2006 [6]. The validation results were summarized in Table 1.

Table 1: Covariate Data Set Accuracies

	CDL Years	Accuarcy	Avg. CDL	Cultivation	Corn/Soy	Wheat	Cotton	Forest	Water	Urban
California	2007 - 2010	Producer	82.82%	98.95%	52.03%	59.50%	66.73%	94.97%	87.96%	93.84%
		User		95.16%	23.93%	21.06%	36.62%	89.17%	91.35%	95.07%
Indiana	2007 - 2010	Producer	94.82%	96.58%	96.74%	39.88%	N/A	93.67%	83.16%	81.83%
		User		89.08%	86.20%	12.71%	N/A	79.36%	73.40%	68.73%
Mississippi	2007 - 2010	Producer	85.79%	84.11%	93.18%	50.65%	67.55%	84.27%	81.56%	84.43%
		User		93.08%	57.46%	23.08%	36.98%	77.68%	79.65%	68.92%
Nebraska	2007 - 2010	Producer	93.06%	98.45%	94.19%	68.44%	N/A	73.67%	73.22%	85.92%
		User		99.63%	83.76%	25.35%	N/A	67.38%	77.23%	64.36%
Pennsylvania	2008 - 2010	Producer	69.74%	74.16%	83.35%	23.94%	N/A	95.29%	80.58%	79.19%
		User		68.48%	53.11%	8.37%	N/A	90.75%	87.01%	78.88%
Washington	2007 - 2010	Producer	90.27%	89.61%	68.01%	90.04%	N/A	98.32%	91.88%	95.36%
		User		88.78%	27.65%	49.93%	N/A	88.29%	90.17%	78.80%

As shown in Table 1, cultivation was identified with high accuracy in 2011 using the 2007- 2010 CDL data for California, Indiana, Mississippi, Nebraska and Washington. The cultivated data sets for those states achieved producer and user accuracies from 84.11 % – 99.63%. The cultivation data set for Pennsylvania performed less well with producer and user accuracies of 74.16% and 68.48%. These results are closely correlated to the accuracies of the original CDLs. The corn/soybean data sets achieved accuracies between 89.08% and 99.63% for the large corn/soybean producing states of Indiana and Nebraska but performed less well in states not dominated by corn and soybean production. However, in states with less corn and soybean acreage, the covariates are less reliable at identifying their planting location in future years. The wheat and cotton covariate data sets were less accurate in predicating wheat and cotton acreage in the test states in 2011. The water, urban and forest data sets achieved generally higher accuracies from 67.38% to 98.32%. They can be relied upon to predict the location of these land cover types in the future, in large part because they change little over time.

Overall, the multi-year cultivation data can be relied upon to predict future cultivation acreage with a high confidence for various applications as the CDLs are highly accurate in crop-intensive areas. The cultivation data sets are currently being used operationally in NASS to stratify new state frames [4].

4. APPLICATIONS

Multi-year CDL based covariate data sets for alfalfa, barley, canola, corn, cotton, grassland, urban, oats, rye, sorghum, soybeans and winter wheat were created for use in building the Oklahoma 2013 Area Sampling Frame. Previously, crop commodity information was derived at the county level from historical NASS county survey estimates. CDL covariate data sets were used to further stratify the NASS Area Frame based on commodity information at the more detailed PSU level.

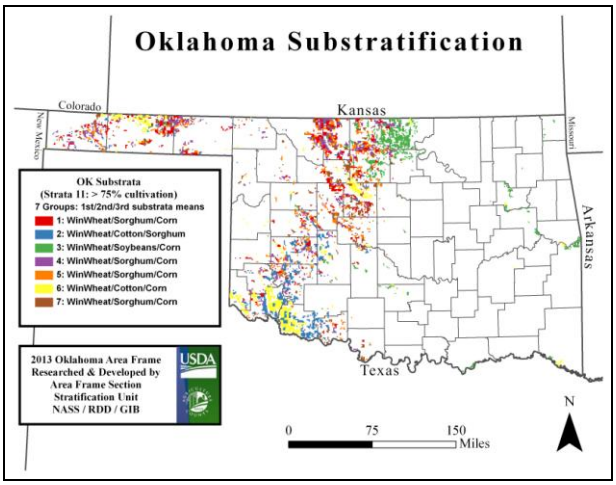

Figure 3. Oklahoma Substratification

Table 2: Stratum 11 Design Effects

Year	Corn	Cotton	Soybeans	Winter Wheat
2012	0.811	0.811	0.773	0.733
2013	0.830	0.683	0.382	0.508

The Oklahoma multi-year cultivated data set was used to help define general strata based on percent cultivated cropland and the covariates defined crop specific substrata as shown in Fig.3. The covariates data sets were used as the inputs to the automated GIS procedure to determine the percent crop at the PSU level [4]. These covariate percentages at the PSU level were used as the input to a simulated annealing procedure to define the optimal substrata clusters based on substrata homogeneity [9].

The design effect of the Oklahoma substratification was evaluated to determine if variances were reduced by using the new CDL covariate based sample design. Table 2 illustrates the design effects calculated for corn, cotton, soybeans and winter wheat in Oklahoma. The 2012 values provide the design effect for the original ASF sample design derived using traditional techniques. The 2013 values

provide the design effect for the new covariate based sample design. Design effects less than 1 indicate an increased precision (reduced variance) in the estimator. Comparing the prior year design effects, the new sample design based on the CDL covariate data shows a reasonable overall improvement in the substratification [9].

New NASS Area Sampling Frames are being created using the PSU level crop data derived from the CDL based covariate data sets. The most intensively cultivated stratum in the new 2014 Arizona, New Mexico, Georgia, South Dakota and North Carolina Area Sampling Frames will be substratified using CDL based covariate data sets information. The data sets will provide the opportunity to objectively, automatically and accurate characterize agricultural content, at the primary sampling unit level.

5. CONCLUSION

This paper presented a new method used to build multi-year CDL cultivation and covariate data sets as well derive crop percentages at the PSU level and the resulting accuracies which reflect their predictive capabilities. The 2011 covariate data sets were developed using 2007-2010 CDL data for California, Indiana, Mississippi, Nebraska, Pennsylvania, and Washington. Areas of cultivation in 2011 were predicted with high accuracy using the multi-year CDL data. Corn and soybeans in states where these crops dominate were also predicted accurately in 2011 with the corn/soybean covariate data sets. Wheat and cotton were predicted with low probability in the test states. The location and extent of forest, water and urban were predicted reliably in the test states. A set of Oklahoma covariate data sets was used to derive percent crop at the PSU level. These crop PSU percentages were used within NASS in 2012 to cluster stratum 11 (greater than 75% cultivation) ASF PSUs into seven substrata based on simulated annealing [9]. The NASS crop specific county estimate data were used to conduct substratification for less cultivated strata. The multi-year CDL covariate data sets provide the unique opportunity to objectively and automatically derive the percentage of specific crop and non crop categories at the ASF PSU level using the CDL based automated stratification method, an objective not previously possible using county estimate statistics.

6. REFERENCES

[1] Benedetti, R., Bee, M., Espa, G., Piersimoni, et al., Agricultural Survey Methods; Chapter 11. *Area Frame Design for Agricultural Surveys*. John Wiley & Sons, Ltd. Published Online: 25 March 2010.

[2] 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.

[3] 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.

[4] Boryan, C. G., and Yang, Z., "A new land cover classification based stratification method for area sampling frame construction," *Proc. in First Intl. Conf. on Agro-Geoinformatics*, Shanghai, China, August 2-4th, 2012.

[5] 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, June, 2012.

[6] Homer, C., J. et. al., "Completion of the 2001 National Land Cover Database for the conterminous United States," *Photogrammetric Eng. and Rem. Sens.* 73 (4):337-341, 2007.

[7] USDA NASS, 2013 <(http://www.nass.usda.gov/research/Cropland/metadata/meta.htm >(last Accessd 28 May 2013)

[8] USDA/FSA, 2013 <http://www.fsa.usda.gov/FSA/apfoapp?area=home&subject=prod&topic=clu-ab> (last accessed 28 May 2013).

[9] Lisic, J. "PSU classification via simulated annealing". Internal NASS report, 5, December, 2012.