DERIVING 2011 CULTIVATED LAND COVER DATA SETS USING USDA NATIONAL AGRICULTURAL STATISTICS SERVICE HISTORIC CROPLAND DATA LAYERS

Claire Boryan[1], Zhengwei Yang[1], Liping Di[2]

[1]National Agriculture Statistics Service, U.S. Department of Agriculture, Fairfax, VA 22030, U.S.A.
[2]George Mason University, Fairfax, VA 22030, U.S.A.

ABSTRACT

This paper describes the method used to derive 30 meter resolution 2011 US cultivated data sets based on multi-year National Agricultural Statistics Service (NASS) Cropland Data Layer (CDL) data. This paper presents different sets of rules (models) to build the cultivated data sets, and a comparison of the resulting cultivated data set accuracies to the accuracies of the original CDL input data. Nine models to create 2011 cultivated data sets for nine US states are tested. Each model provides a set of rules for merging pixels of multi-year (2007-2011) CDL data. The cultivated data accuracy was assessed against in situ 2011 Farm Service Agency (FSA) Common Land Unit (CLU) data. It was found that accuracies were close among the cultivated data generated using the different models. The strongest models for all states achieved overall (producer and user) accuracies greater than 94% for cultivated and non cultivated categories.

Index Terms— cultivated data layer, CDL, land cover, crop mask, multi-year cultivated data layer.

1. INTRODUCTION

Monitoring changes in cropland and specifically in cultivated land (which is prepared by humans and used for crop cultivation, fallow or idle crop land) has become increasingly important due to concerns over climate change and with increased awareness of the impact of agricultural activity on the environment. Within NASS, the cultivated data sets will be used as an input for building Area Sampling Frames (ASF) which are the foundation of the NASS agricultural statistics program and have been used since 1954 for conducting surveys for crop acreage and other agricultural information. The NASS cultivated data sets will be the foundation of a newly developed automated CDL based stratification method [1]. In addition, the accurate identification of cultivated land is critical to many other crop cover related research activities such as crop vegetation condition monitoring. The purpose of this research is to develop and update the state level cultivated data sets based on NASS' annual Cropland Data Layer (CDL) products. The data sets are to be used internally within USDA/NASS and potentially to be disseminated to the public over the NASS CropScape web portal so that scientists, educators, agribusiness and the general public can have access to the most current, highest resolution, accurate characterization of US cultivation available today. The ultimate goal of this study is to determine the best model(s) for creating the state level cultivated data sets and produce not only the 2011 cultivated data sets but produce annual updates in the future.

This paper describes the method used to create nine state level models using the multi-year CDL data to create the 2011 state level raster cultivated data sets. The term "model" here refers to the specific rules used to merge the CDL input data as described in Table 1. It is expected that using multiple years of data rather than simply performing a recode of the 2011 CDLs will result in a more reliable and accurate identification of total cultivation. The multiple years of CDLs for Nebraska, Indiana, Washington, Pennsylvania, Oklahoma, Virginia, Ohio, Georgia, and Mississippi were selected. Nine models were used to produce cultivated data sets for each state and accuracy assessments were run using independent 2011 Farm Service Agency (FSA) Common Land Unit (CLU) data as validation to compare the models.

2. BACKGROUND AND STUDY DATA

Global and conterminous US land cover products have been used since the mid 1980s to identify cultivation and specific types of vegetation [2, 3, 4, 5]. However, the spatial resolution of these global products is too coarse and the accuracy is not sufficient for many applications. The land cover data products for 48 conterminous states in the US such as the National Land Cover Data (NLCD) set and CDL [6, 7] have higher spatial resolution. The NLCD was first produced in 1992 and updated for years 1997, 2001. The NLCD 2006 product has 16 land cover categories and a 30 meter spatial resolution. However, its cropland cover is not as accurate as NASS' CDL products. The USDA NASS' US CDL products are produced at 30 – 56 meter resolution. The CDL products accurately identify and geolocate field crops. Total crop map accuracies for the historic CDLs range from 85% to 95% for the major crop categories [7]. The CDL data are publically available from NASS' online geospatial

application – CropScape [8]. The NASS CDLs are unique because they are updated yearly with the primary purpose of crop identification. Approximately 110 different types of crops across the US are identified in the NASS CDLs. Furthermore, the NASS CDLs are created using USDA's Farm Service Agency (FSA) Common Land Unit (CLU) data, a voluminous source of in situ crop training data.

In this study, multi-year (2007 - 2011) or 2008-2011 CDLs for the first time were utilized as the inputs to create highly accurate 30 meter cultivated data sets for nine pilot states (Nebraska, Indiana, Washington, Pennsylvania, Oklahoma, Virginia, Ohio, Georgia, and Mississippi) in the US. These states were selected to reflect the range of agriculture grown in the United States. When possible, all five 2007-2011 CDLs were used as inputs to the cultivation models to generate the cultivated data layer. For Virginia, Georgia and Pennsylvania only 2008 - 2011 CDLs were available. The 2007 - 2009 CDLs have a 56 meters spatial resolution and the 2010 and 2011 CDLs have a 30 meters spatial resolution. Some of the 2008 CDLs were processed in 2011 and are 30 meters.

For validation purposes independent 2011 FSA CLU and NLCD sample point data were used to assess the accuracy of the cultivated data sets produced using the eight or nine different models for each state. The FSA data provides crop information and the NLCD provides information on non agricultural categories. These are the same validation data sets used to assess the accuracy of the 2011 CDLs, just recoded to cultivated and non cultivated categories [7].

3. CULTIVATED MASKING METHODOLOGY

The procedure for generating a cultivated data set based on NASS CDL raster data is straightforward. All pixels of the original CDLs and validation data sets were re-coded from their original categories into cultivated and non cultivated categories. In recoding, a cultivated category (including all crop types with the exception of non alfalfa hay) was identified with a "1" and a non cultivated category was identified with a "0". After recoding, the individual CDLs were combined at the pixel level based on model rules as shown in Table 1 to create state level 2007–2011 or 2008-2011 data sets. Table 1 illustrates 9 different models which define different rules used to merge the recoded individual cultivated layers. For example, in Model 5, the merged cultivated data set retained those pixels that were categorized to a cultivated crop at least three times in the original CDL inputs. In the same model used to recode and merge the data sets, the 2007-2009 CDLs were resampled to 30 meters using nearest neighbor interpolation. The output file included categories 1 – 5 indicating whether a pixel was categorized as a cultivated crop 1, 2, 3, 4, or 5 times in the original CDL inputs during the period from 2007 – 2011. Models 1 - 9 were used to create cultivated data sets for all nine states in this investigation.

Table 1: Model (Rule Sets) for Building 2011 State Level Cultivated Data Set

Model No.	Model Name	Description of Models Pixels
1	Cultivated in 2011	All pixels categorized to cultivated crop in the 2011 CDL only are retained to build the 2011 cultivated data set. No multiyear data are used
2	Cultivated in 2009-2011	All pixels categorized to cultivated crop in the 2009-2011 CDL are retained. There are no 2007 – 2008 data.
3	Cultivated 1+ Years	All pixels ever categorized to a cultivated crop in the all available years are retained.
4	Cultivated 2+ Years	Pixels categorized to a cultivated crop at least two times in the original CDL inputs are retained.
5	Cultivated 3+ Years	Pixels categorized to a cultivated crop at least three times in the original CDL inputs are retained.
6	Cultivated 4+ Years	Pixels categorized to a cultivated crop at least four times in the original CDL inputs are retained
7	Cultivated 5 Years	Pixels categorized to a cultivated crop in all available years in the original CDL inputs are retained.
8	Cult in 2011 or Cult 2+ Years	Pixels categorized to a cultivated crop in the 2011 CDL as well as pixels categorized to a cultivated crop at least two times in the original CDL inputs are retained.
9	Cult in 2011 or Cult 3+ Years	Pixels categorized to a cultivated crop in the 2011 CDL as well as pixels categorized to a cultivated crop at least two times in the original CDL inputs are retained.

4. RESULTS AND DISCUSSION

A final state cultivated data product is exemplified by Figure 2, which illustrates the 2011 Nebraska cultivated data set. This product was created by merging the original 2007-2011 Nebraska Cropland Data Layers. The land represented by the green pixels was identified as a cultivated (crop) category at least once during the five year period. The 2011 Nebraska Cultivated Data Set has a producer accuracy of 99.20% (cultivation) and 90.94% (non cultivation); a user accuracy of 99.58% (cultivation) and 84.05% (non cultivation) and an overall accuracy (producer and user combined) of 98.89%.

Similar products were created for all nine states. Table 2 summarizes the accuracy results derived from comparing the Model 1-9 cultivated data sets to the state specific 2011 FSA CLU and NLCD validation data. As shown in the Table 2, Models 1 – 9 correspond to different CDL input combinations. The average overall CDL accuracies were listed for comparing with Model 1 – 9 cultivated data set accuracies. Among nine test states, four states were not applicable to Model 7 because there were only four years of

input data. The highlighted green cells in Table 2 identify the top two model cultivation accuracies for each state. Since the original CDL input data are accurate in the identification and differentiation of specific crop types, this translates into high accuracies of the resulting cultivated data sets. The highest accuracies achieved by all state models were between 75.75% - 99.0% with most achieving accuracies above 90%. The lowest accuracies (75.75% - 91.42%) were achieved using Model 7 which required that only pixels categorized to a cultivated crop in all years (5+) be retained to build the 2011 cultivated data set. The high accuracies achieved by the NASS 2011 cultivated data sets are, as previously mentioned, due in large part to the high accuracies of the original (2007-2011) CDL input data.

Upon review of the state level cultivated data set accuracy results (Table 2), no single model was able to achieve the highest accuracy for all states. Nebraska (98.84%), Indiana (99.0%), Ohio (98.40%), and Oklahoma (98.77%) achieved the highest overall accuracies using Model 3 in which all pixels ever categorized to a cultivated crop in the 2007-2011 CDLs were retained to build the more complete 2011 cultivated data set. Model 2, which included all pixels categorized to a cultivated crop from 2009-2011, achieved the second highest accuracies for these states. These results indicate, albeit by a small margin, that using five years of CDL data results in higher accuracies then using three years of CDL input data.

The two highest model accuracies for Pennsylvania were achieved using Model 8 (94.93%) and Model 4 (94.78%). Both of these models filter out one year of data and require that pixels be categorized to a cultivated category two or more times to be retained. The two highest model accuracies for Virginia were Model 8 (96.20%) and Model 9 (96.22%) which filter out either one or two years of data but always include the 2011 cultivated information. The two highest model accuracies for Mississippi were Models 4 (96.56%) and Model 5 (96.49%) which again filter out between one and two years of data. The models for two states, Washington (97.74%) and Georgia (97.07%) achieved the highest accuracies using the 2011 CDL alone. The Model 9 results were very close because they also include the 2011 cultivated data.

The range in accuracies achieved using Models 1-9 for all states is relatively small with many ranging from the high 80% to high 90% in overall accuracy. Two conclusions can be drawn. First, using only the single year 2011 CDLs to create the 2011 cultivated data sets (Model 1) did not achieve the highest accuracy in the majority (7 out of 9) of states. Second, requiring a pixel to be identified as cultivation in all available years (Model 7) is too restrictive and achieved the lowest accuracies.

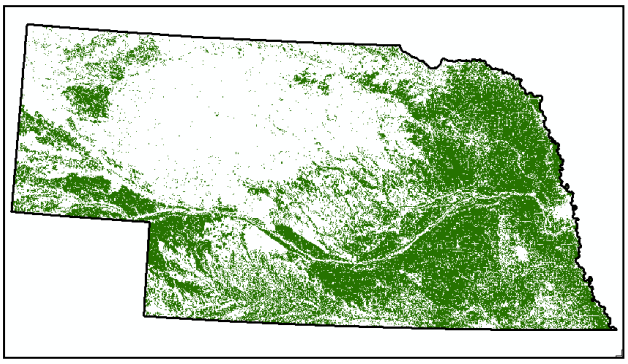

Figure 2: 2011 Nebraska Cultivated Data Set. Green pixels identify cultivation. White pixels identify non cultivation.

Two predominant model groups achieved the highest accuracies. First, Models 2/3 which retained all cultivated pixels in all years or in the three years from 2009-2011 achieved the highest accuracies in states that had the higher original overall CDL accuracies (Table 2). Second, Model 8/9 which retained cultivated pixels identified at least two or three times and included all cultivated pixels in the 2011 CDL achieved the highest accuracies in states that had lower average overall CDL accuracies.

To understand why Models 2 and 3 were most accurate for Nebraska (98.84%), Indiana (99.0%), Ohio (98.40%), and Oklahoma (98.77%), it is useful to examine the average crop accuracies (across all available years) for the original CDL products (Table 2). Indiana, Nebraska and Ohio had the highest average crop (CDL) accuracies of the nine states. Consequently, it can be assumed that because the original CDLs were highly accurate, then all available input data can be relied upon and used in building the cultivated data set. Oklahoma is interesting in that it ranges in the mid level of accuracy for the CDLs and Model 3 still achieved the highest accuracy.

To understand why Model 8/9 were more accurate for Pennsylvania (94.93%), Virginia (96.20%), and Georgia (96.98%), we need to remember that this process filters out classification errors in the original CDLs and produces a more accurate 2011 cultivated data sets. It is preferable to use this type of model that filters out classification errors when the original CDLs have lower accuracy. The CDL average crop accuracy for Virginia was only 71.83%, Pennsylvania was 70.36% and Georgia was 78.86% while the Indiana, Nebraska and Ohio CDL average crop accuracies were all above 91.0% (Table 2).

Based on this research, it can be assumed that building the NASS 2011 cultivated data sets for all 48 conterminous states should follow these basic guidelines. First, the 2011 NASS cultivated data sets should be created at the state level since there is a wide range of CDL average crop accuracies from a low of 58.83% for Alabama and a high of 96.37% for Illinois.

Table 2: Cultivated Data Set Accuracy Results for Models 1 -9

	CDL Years	Average CDL Accuracy	Model 1 Cultivated in 2011	Model 2 Cultivated in 2009 - 2011	Model 3 Cultivated 1+ Years	Model 4 Cultivated 2+ Years	Model 5 Cultivated 3+ Years	Model 6 Cultivated 4+ Years	Model 7 Cultivated 5 Years	Model 8 Cult. in 2011 or Cult. 2+ Years	Model 9 Cult. in 2011 or Cult. 3+ Years
Indiana	2007 - 2011	94.74%	97.66%	**98.93%**	**99.00%**	98.05%	95.04%	92.78%	88.02%	98.66%	98.43%
Nebraska	2007 - 2011	93.43%	97.08%	**98.66%**	**98.84%**	97.70%	95.33%	91.78%	82.76%	98.36%	97.93%
Ohio	2007 - 2011	91.17%	96.78%	**98.19%**	**98.40%**	96.70%	93.12%	89.93%	83.22%	97.78%	97.44%
Washington	2008 - 2011	90.82%	**97.74%**	97.24%	96.67%	96.38%	93.12%	87.22%	N/A	97.49%	**97.71%**
Mississippi	2007 - 2011	83.84%	96.04%	94.70%	94.22%	**96.56%**	**96.49%**	95.05%	91.15%	96.20%	96.34%
Oklahoma	2007 - 2011	83.54%	95.09%	**98.40%**	**98.77%**	97.28%	93.66%	88.74%	76.22%	98.03%	97.24%
Georgia	2008 - 2011	78.86%	**97.07%**	94.96%	94.32%	94.36%	89.40%	75.75%	N/A	96.58%	**96.98%**
Virginia	2008 - 2011	71.83%	95.83%	95.56%	95.07%	95.89%	94.71%	91.42%	N/A	**96.20%**	**96.22%**
Pennsylvania	2008 - 2011	70.36%	94.06%	94.63%	93.58%	**94.78%**	92.91%	88.98%	N/A	**94.93%**	94.75%

For states that have average CDL crop accuracies above 90% then Model 3 should be used to build the cultivated data set. The 2011 NASS cultivated data sets for Nebraska, Indiana, Ohio, Oklahoma, Illinois, Iowa and Minnesota should be built using Model 3 based on the accuracy statistics posted on the NASS Research and Division web site. These states have a very high percentage of cultivation, and the highest percentage of FSA CLU training data with which to produce the CDLs so they tend to have the highest CDL accuracies over time. Cultivated data sets that are created for states that have lower original CDL average crop accuracies (below 90%) should be created using Model 9 which allows for classification errors in the original CDL input data to be filtered (2+years) but retains the cultivated information from the 2011 CDLs.

The 2011 NASS cultivated data sets should be generated at the state level using either Model 3 or 9 based on state specific CDL average crop accuracies and then merged to create a NASS 2011 Cultivated Data Set which includes all 48 conterminous states. Further, five years of CDL data are optimal for use in all models in future years to take advantage of continual improvements in NASS CDL processing techniques such as improvements in satellite data, spatial resolution, training data or classification procedures.

5. CONCLUSION

This paper describes the method used to develop new 2011 state level cultivated data sets, the resulting accuracies, the correlation between the cultivated data set accuracy statistics and the accuracies of the original CDL input data, and suggested guidelines for building a 2011 NASS cultivated data set for all 48 conterminous states. The strongest models for all states achieved overall (producer and user) accuracies greater than 94% for cultivated and non cultivated categories. Overall the differences in model accuracies were very small at the state level because the CDL input data were consistent over time. However, at the local level these differences may be significant. The accuracies of the cultivated data sets are highly dependent on the original CDL accuracies. Cultivated data sets created using multi-year NASS Cropland Data Layer products provide the unique opportunity to update and reflect changes in cultivated land for the 48 conterminous states in the US on a yearly basis which is not possible with other global or US land cover data sets available today. Moreover, the method used to develop the US cultivated data sets is straightforward.

6. REFERENCES

[1] 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.

[2] Matthews, E., "Global vegetation and land use: New high-resolution data bases for climate studies." *J. Clim. Appl. Meteor*. 22, 474-487, 1983.

[3] Loveland, T.R., et al., "Development of a Global Land Cover Characteristics Database and IGBP DISCover from 1-km AVHRR Data." *International Journal of Remote Sensing*, v. 21, no. 6/7, p. 1,303-1,330. 2000.

[4] Hansen, M. C., DeFries, R. S., Townshend, J. R. G., and Sohlberg, R., "Global land cover classification at 1 km spatial resolution using a classification tree approach." *International Journal of Remote Sensing*, 21, 1331–1364, 2000.

[5] Friedl, M.A., Brodley, C.E. and A.H. Strahler, "Maximizing land cover classification accuracies at continental to global scales," *IEEE Trans. on Geosci. and Rem. Sens.*, vol.37, pp. 969-977, 1999.

[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] 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, 2011.

[8] 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," *Comput. and Elect. in Agric.* Vol. 84, June, pp. 111–123, 2012.