Crop Specific Covariate Data based on the NASS Cropland Data Layer for Area Frame Stratification

Claire G. Boryan, Geographer
USDA\NASS\Research and Development Division
Fairfax, VA

“... providing timely, accurate, and useful statistics in service to U.S. agriculture.”
Background

- NASS provides timely, accurate, and useful statistics in service to U.S. agriculture
Applications of the Cropland Data Layer within NASS

• Crop Acreage Estimation (1997-2012)
• Area Sampling Frame Stratification (2010-2013)
• Area Sampling Frame Substratification (2013)

“... providing timely, accurate, and useful statistics in service to U.S. agriculture.”
What are Area Sampling Frames?

- NASS Area Sampling Frames been used as the **primary tool** to conduct agricultural surveys since 1954.

- The NASS Area Sampling Frames are the basis for the annual **June Area Survey** in which approximately 11,000 segments are enumerated in early June to collect crop acreage and other agricultural information.

- NASS ASFs are based on a **stratification of land cover** in the U.S. defined by percent cultivated cropland.
What is the stratification?

• The word “stratify” comes from the Latin words meaning “to make layers”; we divide the population into subpopulations, called strata.

• In statistics, the goal is to make the strata as homogeneous as possible.

• Stratified sampling generally gives more precise (lower variance) estimates for population means and totals than simple random sampling alone.
How Is Stratification Performed at NASS?

• It has been conducted by Area Frame staff since 1954 using **visual interpretation** of aerial photography, and later moderate resolution Landsat TM data.

• The NASS Cropland Data Layer products have been used in recent years to aid in the visual interpretation process.

• In the past two years **Cropland Data Layer (CDL) automated stratification** has begun to be implemented.

“... providing timely, accurate, and useful statistics in service to U.S. agriculture.”
What is the Cropland Data Layer (CDL)?

The Cropland Data Layer product is a raster-formatted, crop specific, land cover data set used to generate in-season acreage estimates for 19 crops in 41 states.

- Corn
- Soybeans
- Winter Wheat
- Cotton
- Rice
- Alfalfa

Released Jan. 31, 2013
National 30m Product
High accuracy for major crops
2012 Cropland Data Layer Inputs

Satellite Imagery – Landsat

Satellite Imagery – Deimos & UK2

Farm Service Agency: Common Land Unit

2006 NLCD & Derivative products
National Cropland Data Layer Legend

2011 Continental United States Land Cover Categories (by decreasing acreage)

Agriculture

- Pasture/Grass
- Corn
- Soybeans
- Winter Wheat
- Fallow/Idle Cropland
- Other Hay/Non Alfalfa
- Alfalfa
- Cotton
- Spring Wheat
- Sorghum
- Dbl Crop WinWht/Soybeans
- Rice
- Barley
- Oranges
- Cuts
- Sunflower
- Dry Beans
- Peanuts
- Durum Wheat
- Sugarcane
- Potatoes
- Canola
- Sugarcane
- Almonds
- Soy/Grass Seed
- Grapes
- Oat
- Rye
- Peas
- Millet
- Walnuts
- Lentils
- Pecans
- Dbl Crop WinWht/Cotton
- Dbl Crop WinWht/Sorghum
- Sweet Corn
- Aquaculture
- Clover/Wildflowers
- Other Crops
- Dbl Crop WinWht/Com
- Cherries
- Triticale
- Citrus
- Sunflower
- Pistachios
- Blueberries
- Christmas Trees
- Dbl Crop Barley/Soybeans
- Tomatoes
- Onions
- Flaxseed
- Dbl Crop Dats/Corn
- Pop or Corn Corn
- Herbs
- Misc Vgs & Fruits
- Olives
- Other Tree Crops
- Dbl Crop Corn/Soybeans
- Sweet Potatoes
- Peaches
- Cranberries
- Tobacco
- Cattail/owes
- Prunes
- Dbl Crop Barley/Com
- Dbl Crop Soybeans/Cotton
- Peas
- Lettuce
- Dbl Crop Barley/Sorghum
- Dbl Crop Durum Wht/Sorghum
- Watermelons
- Switchgrass
- Asparagus
- Carrots
- Strawberries
- Pumpkins
- Squash
- Cabbage
- Peppers
- Dbl Crop Soybeans/Oats
- Hops
- Mint
- Pomegranates

Non-Agriculture

- Forest
- Shrubland
- Developed
- Wetlands
- Water
- Barren
- Perennial Ice/Snow
CropScape

http://nassgeodata.gmu.edu/CropScape

Interactive CDL visualization/exploration/dissemination portal
Research is being conducted to develop new methods and data sets to utilize the CDL in the construction of the NASS Area Sampling Frame (ASF).
The NASS 2012 Cultivated Layer created from 2008-2012 CDL data

Validation based on independent, 2012, FSA CLU and NLCD, 2006 data

<table>
<thead>
<tr>
<th>Class Name</th>
<th>Code</th>
<th>Producer Accuracy</th>
<th>Omission Error</th>
<th>Kappa</th>
<th>User Accuracy</th>
<th>Commission Error</th>
<th>Conditional Kappa</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-Cultivated</td>
<td>1</td>
<td>96.6%</td>
<td>3.4%</td>
<td>0.940</td>
<td>97.2%</td>
<td>2.8%</td>
<td>0.951</td>
</tr>
<tr>
<td>Cultivated</td>
<td>2</td>
<td>97.9%</td>
<td>2.1%</td>
<td>0.951</td>
<td>97.4%</td>
<td>2.6%</td>
<td>0.940</td>
</tr>
</tbody>
</table>

Applications of the Cropland Data Layer within NASS

- Crop Acreage Estimation (1997-2012)
- Area Sampling Frame Stratification (2010-2013)
- Area Sampling Frame Substratification (2013)

“... providing timely, accurate, and useful statistics in service to U.S. agriculture.”
Automated Stratification of the NASS Area Sampling Frame based on the CDL

Primary Sampling Units with CDL percent cultivation

Primary Sampling Units with CDL percent cultivation, overlaying a 2010 CDL image product
CDL based stratification of a NASS Area Sampling Frame (ASF)
Stratum & Percent Cultivation

Accuracy based on 2010 June Area Survey in situ data
New state Area Sampling frames created using the CDL automated stratification method

- Arizona - 2014
- Georgia - 2014
- New Mexico - 2014
- Oklahoma - 2013
- South Dakota - 2014
- North Carolina - 2014

“... providing timely, accurate, and useful statistics in service to U.S. agriculture.”
The Oklahoma Area Sampling Frames  (2010 and 2013). Stratum 11 (>75% cultivated) was overestimated in the 2010 ASF which was created using the traditional method and updated to more accurately reflect conditions in the 2013 ASF using the CDL automated stratification method.  (Graphic courtesy of Kevin Hunt - AF Section -NASS)
New state frames are being built at reduced cost with improved objectivity, efficiency and accuracy.

- The 2010 Oklahoma frame (traditional method) - 4552 employee hours

- The 2013 Oklahoma frame (CDL automated method) – 1980 employee hours

- Predictive accuracy of the 2013 Oklahoma frame – 78% of segments meeting stratum definition vs. 34% for the 2010 Oklahoma frame (based on 2012 JAS data)

“... providing timely, accurate, and useful statistics in service to U.S. agriculture.”
Applications of the Cropland Data Layer within NASS

• Crop Acreage Estimation (1997-2012)
• Area Sampling Frame Stratification (2010-2013)
• Area Sampling Frame Substratification (2013)

“... providing timely, accurate, and useful statistics in service to U.S. agriculture.”
CDL (2007-2010) covariates at the 2011 JAS segment level

CA11 Cropland Data Layer with JAS segments

Multi Year (2007-2010) cultivated data set with JAS segments (percent cultivation calculated)
CDL (2007-2010) covariates at the 2011 JAS segment level

Multi-year (2007-2010) cotton data set

Multi Year (2007-2010) corn/soy data set
## CDL Covariate Predictive Accuracy

<table>
<thead>
<tr>
<th></th>
<th>CDL Years</th>
<th>Accuracy</th>
<th>Avg. CDL</th>
<th>Cultivation</th>
<th>Corn/Soy</th>
<th>Wheat</th>
<th>Cotton</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>California</strong></td>
<td>2007 - 2010</td>
<td>Producer</td>
<td>82.82%</td>
<td>98.95%</td>
<td>52.03%</td>
<td>59.50%</td>
<td>66.73%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>User</td>
<td></td>
<td>95.16%</td>
<td>23.93%</td>
<td>21.06%</td>
<td>36.62%</td>
</tr>
<tr>
<td><strong>Indiana</strong></td>
<td>2007 - 2010</td>
<td>Producer</td>
<td>94.82%</td>
<td>96.58%</td>
<td>96.74%</td>
<td>39.88%</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td></td>
<td>User</td>
<td></td>
<td>89.08%</td>
<td>86.20%</td>
<td>12.71%</td>
<td>N/A</td>
</tr>
<tr>
<td><strong>Mississippi</strong></td>
<td>2007 - 2010</td>
<td>Producer</td>
<td>85.79%</td>
<td>84.11%</td>
<td>93.18%</td>
<td>50.65%</td>
<td>67.55%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>User</td>
<td></td>
<td>93.08%</td>
<td>57.46%</td>
<td>23.08%</td>
<td>36.98%</td>
</tr>
<tr>
<td><strong>Nebraska</strong></td>
<td>2007 - 2010</td>
<td>Producer</td>
<td>93.06%</td>
<td>98.45%</td>
<td>94.19%</td>
<td>68.44%</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td></td>
<td>User</td>
<td></td>
<td>99.63%</td>
<td>83.76%</td>
<td>25.35%</td>
<td>N/A</td>
</tr>
<tr>
<td><strong>Pennsylvania</strong></td>
<td>2008 - 2010</td>
<td>Producer</td>
<td>69.74%</td>
<td>74.16%</td>
<td>83.35%</td>
<td>23.94%</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td></td>
<td>User</td>
<td></td>
<td>68.48%</td>
<td>53.11%</td>
<td>8.37%</td>
<td>N/A</td>
</tr>
<tr>
<td><strong>Washington</strong></td>
<td>2007 - 2010</td>
<td>Producer</td>
<td>90.27%</td>
<td>89.61%</td>
<td>68.01%</td>
<td>90.04%</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td></td>
<td>User</td>
<td></td>
<td>88.78%</td>
<td>27.65%</td>
<td>49.93%</td>
<td>N/A</td>
</tr>
</tbody>
</table>

Validation: 2011 Farm Service Agency Common Land Unit/ NLCD 2006 – cultivation
2011 Cropland Data Layers - corn/soy, wheat, cotton
Applications of CDL Covariate Data within NASS

(courtesy Jonathan Lisic and Kevin Hunt)

• In the past, commodity information was derived at the county level to infer the agricultural makeup for an entire county.

• CDL covariate data sets provide the opportunity to automatically substratify the NASS Area Frame based on commodity information at the Primary Sampling Unit level.

“... providing timely, accurate, and useful statistics in service to U.S. agriculture.”
Oklahoma 2013 substratification

• Problem: Find the assignment of $N_h$ sampling units to $H$ strata that minimizes the sample size

$$n = \sum_{h=1}^{H} n_h$$

• Subject To

$$T_j \geq \sum_{h=1}^{H} \frac{N_h^2 S_{h,j}^2}{n_h}$$

• Where $T_j$ is a target variance for commodity $j$ for stratum $h$ in \{1,...,H\}

$$S_{h,j}^2 = (N_h - 1)^{-1} \sum_{i=1}^{N_h} (x_{i,j} - \bar{x}_j)_h^2$$
Oklahoma Substratification

OK Substrata
(Strata 11: > 75% cultivation)
7 Groups: 1st/2nd/3rd substrata means
- 1: WinWheat/Sorghum/Corn
- 2: WinWheat/Cotton/Sorghum
- 3: WinWheat/Soybeans/Corn
- 4: WinWheat/Sorghum/Corn
- 5: WinWheat/Sorghum/Corn
- 6: WinWheat/Cotton/Corn
- 7: WinWheat/Sorghum/Corn

2013 Oklahoma Area Frame
Researched & Developed by
Area Frame Section
Stratification Unit
NASS / RDD / GIB
Design Effects

• “The design effect provides a measure of the precision gained or lost by use of the more complicated design instead of Simple Random Sampling (SRS)” (Lohr, S., 2010)

• The design effect is equal to the
  • variance (estimator from sampling plan) divided by
  • variance (estimator from a SRS in stratum 11 with the same # of observation units)

• Design effect values less than 1 indicate an increased precision (reduced variance) in the estimator

“... providing timely, accurate, and useful statistics in service to U.S. agriculture.”
Design Effects

• Comparing prior year design effects using CDL covariate data shows a reasonable overall improvement in substratification efficiency.

<table>
<thead>
<tr>
<th>Year</th>
<th>Corn</th>
<th>Cotton</th>
<th>Soybeans</th>
<th>Winter Wheat</th>
</tr>
</thead>
<tbody>
<tr>
<td>2012</td>
<td>0.811</td>
<td>0.811</td>
<td>0.773</td>
<td>0.733</td>
</tr>
<tr>
<td>2013</td>
<td>0.830</td>
<td>0.683</td>
<td>0.382</td>
<td>0.508</td>
</tr>
</tbody>
</table>

“... providing timely, accurate, and useful statistics in service to U.S. agriculture.”
Conclusion

• The strength of the NASS Cropland Data Layer (CDL) product and CDL based stratification method is the objective and consistent identification of cultivated cropland.

• Utilizing the CDL data for Area Frame stratification and sub-stratification will improve the efficiency, reduce the cost and improve the precision of the June Agricultural Survey estimates.

“... providing timely, accurate, and useful statistics in service to U.S. agriculture.”
Thank you

Questions?

Claire G. Boryan, Geographer
USDA/NASS/RDD
Feb 22, 2012

“... providing timely, accurate, and useful statistics in service to U.S. agriculture.”