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

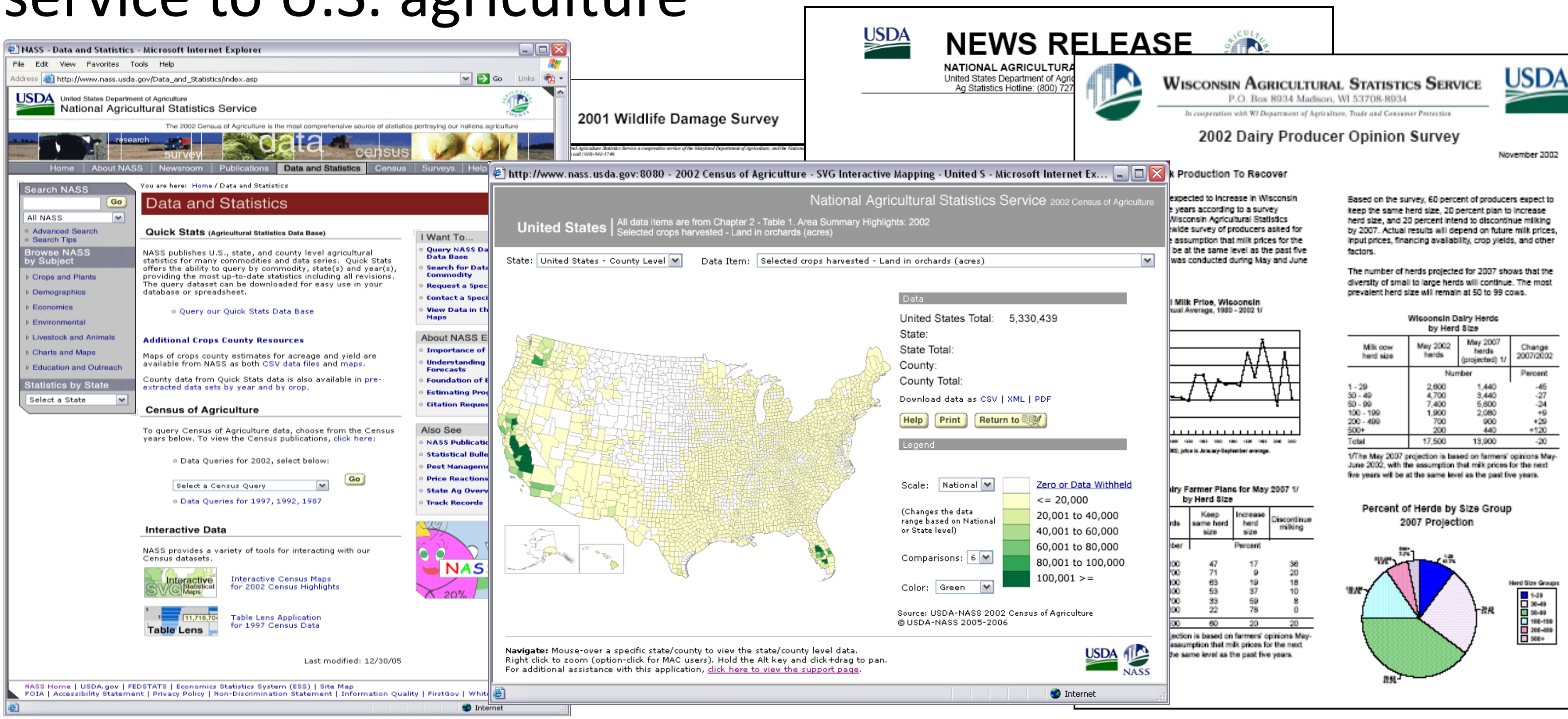
Claire G. Boryan, Zhengwei Yang
USDA\NASS\Research and Development Division
Fairfax, VA
Claire.Boryan@nass.usda.gov
Zhengwei.Yang@nass.usda.gov

Outline

- Background
- Covariate Stratification based on CDL
- Results
- Ultimate Effect Assessment
- Working in Progress
- Conclusions

Background

- NASS provides timely, accurate, and useful statistics in service to U.S. agriculture



How Does Agricultural Statistics Collected at NASS?

- Agriculture Census every five years
- Estimates from Remote Sensing
- Agricultural surveys
  - Estimates from samples based on NASS area sampling frames(ASFs) and list frames

What Is An Area Sampling Frames?

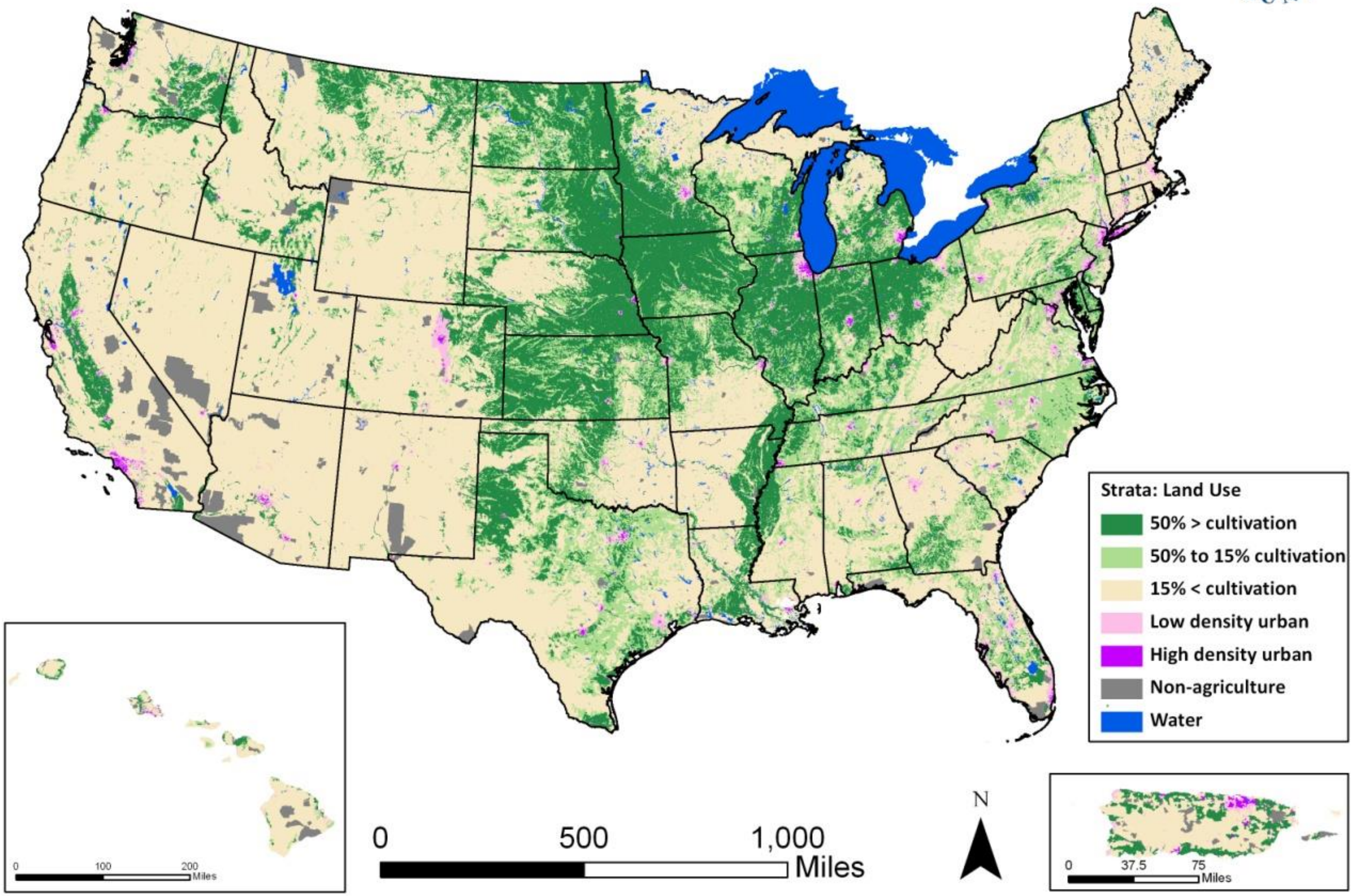
- An area sampling frame is a collection of segmented land parcels for the area of interest, such as a state. A land parcel can be defined by its attributes, such as ownership, land usage, land cover, etc.

- NASS ASFs are based on a **stratification of land cover** in the U.S. defined by percent cultivated cropland, i.e. all land parcels are classified different land cover categories!

- NASS Area Sampling Frames have 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.

National Agricultural Statistics Service
Land Use Area Frame



What Is A Covariate?

- (From Wikipedia) "In statistics, a covariate is a variable that is possibly predictive of the outcome under study. A covariate may be of direct interest or it may be a confounding or interacting variable."
- For NASS, covariates are variables that may be predictive of where crops will be grown in the future.
- They are derived based on the Area Sampling Frames

Why Covariates?

- To further improve crop estimates, the ASF Primary Sampling Units (PSUs) have to be substratified based on crop specific information rather than solely on percent cultivation, i.e. the crop specific covariates have to be derived.
- Covariates may be predictive of where crops will be grown in the future.
- The ASFs are built for future use.
- A covariate is a crop specific variable. It can be used to improve area sampling design and ultimately survey estimation for individual major crops.

How Is Stratification Performed at NASS?

- Stratification goal – make the strata as homogeneous as possible so that the Stratified sampling generally gives more precise (lower variance) estimates for population means and totals than simple random sampling alone.

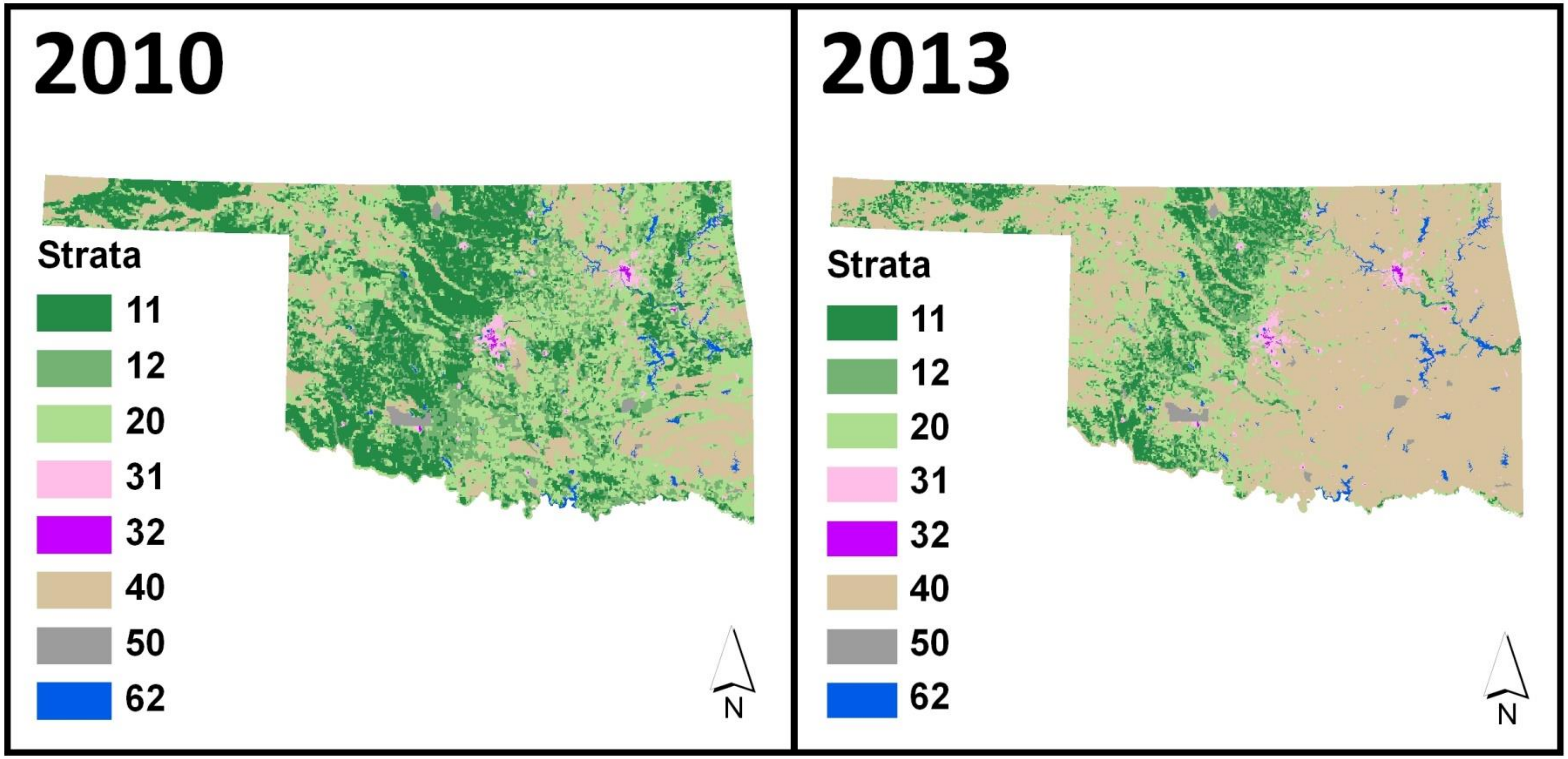
- Stratification has been conducted by Area Frame staff since 1954 using **visual interpretation** of aerial photography, and later moderate resolution Landsat TM data.

- Digital technology (computerized) was introduced since 1993.

- 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) - based automated stratification** has begun to be implemented.

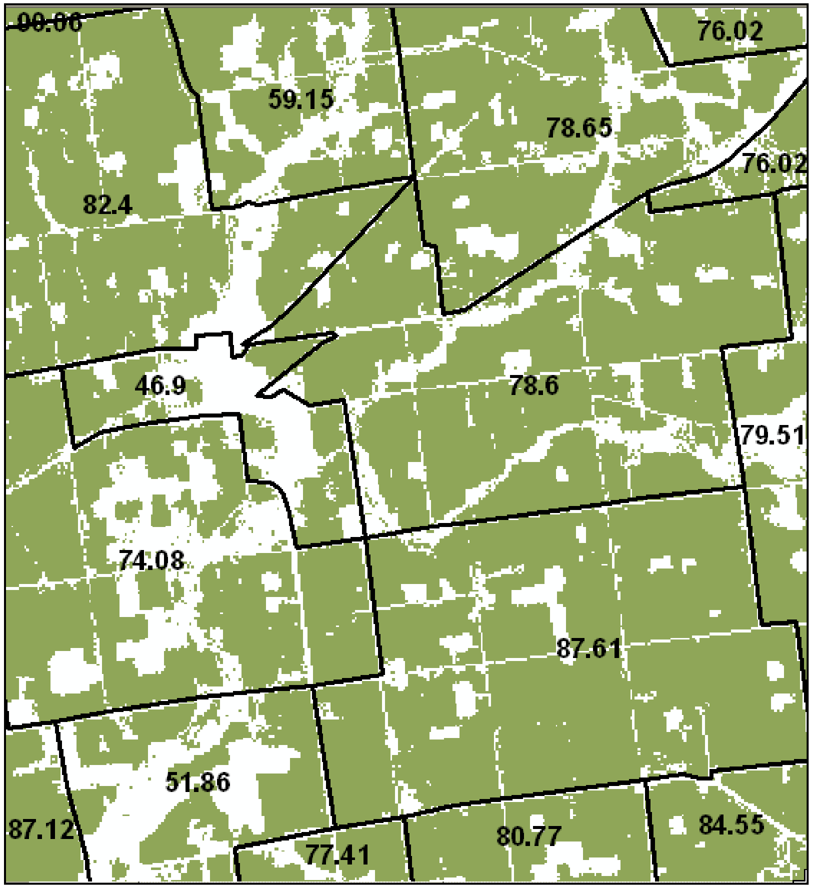
CDL Automated Stratification in NASS Operations



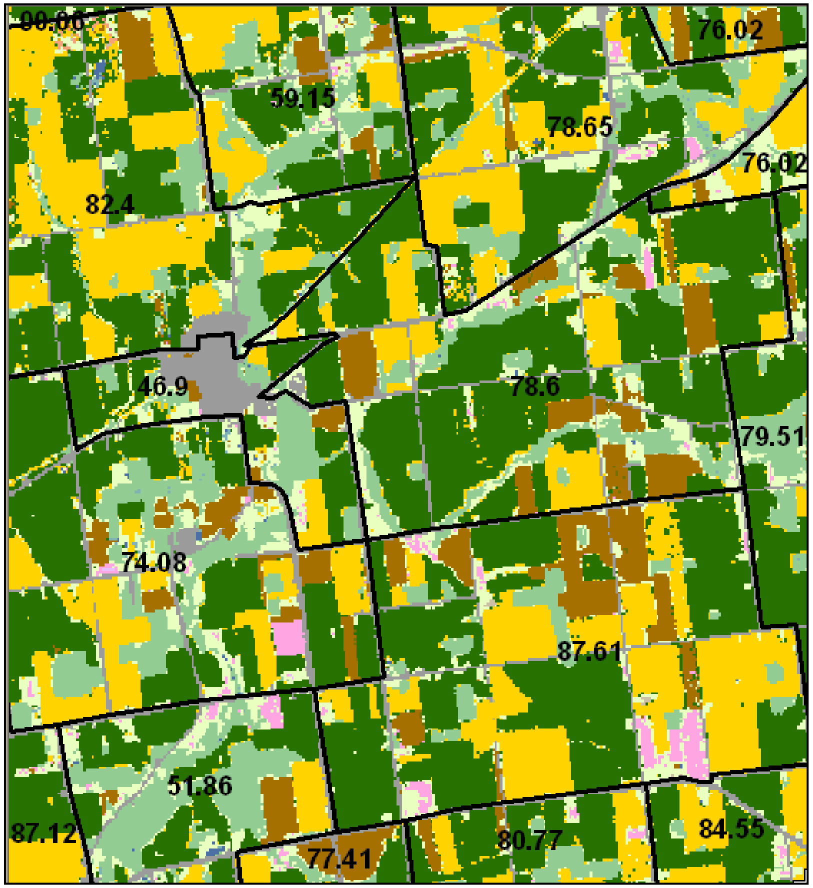
The Oklahoma Area Sampling Frames (2010 and 2013). Stratum 11 (>75% cultivated) was overestimated in the 2010 ASF which was created using the traditional method and updated to more accurately reflect conditions in the 2013 ASF using the CDL automated stratification method. (*Graphic courtesy of Kevin Hunt - AF Section -NASS*)

Automated Covariate Stratification

Automated Stratification of the NASS Area Sampling Frame based on the CDL



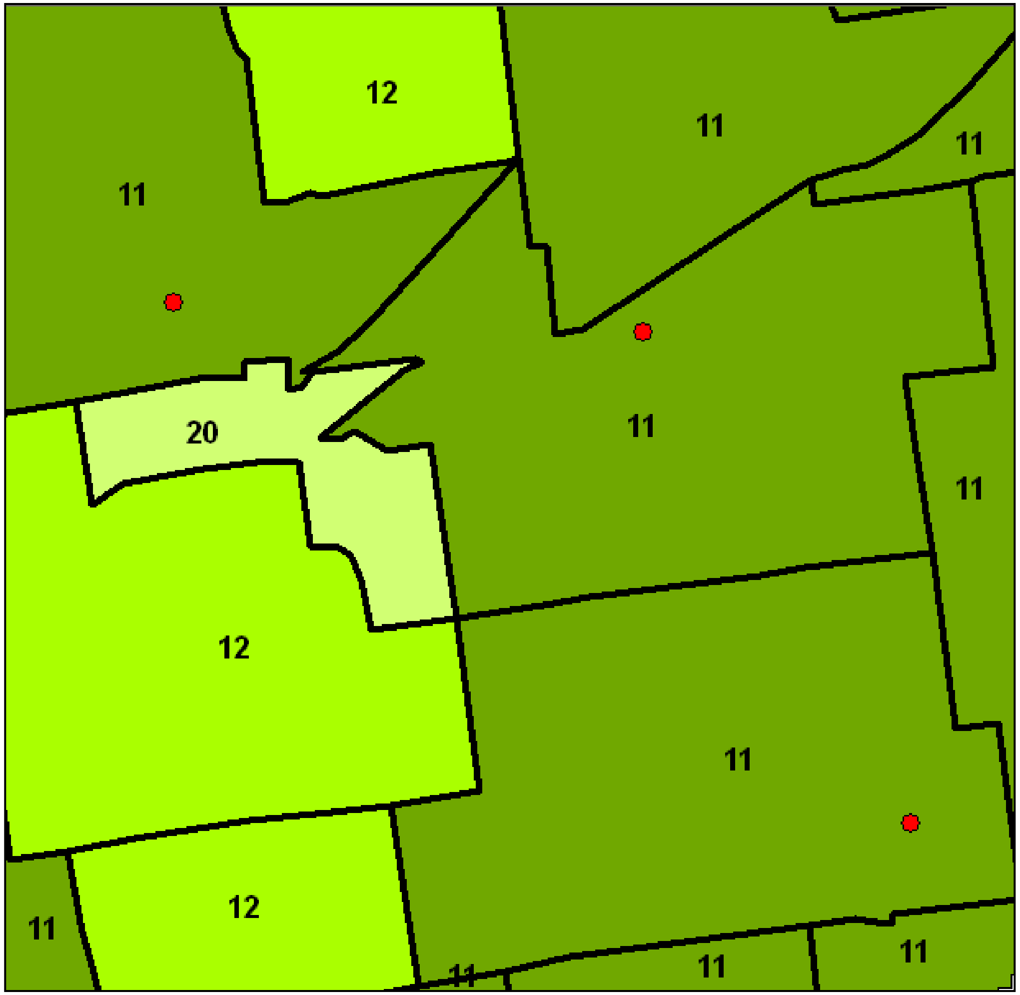
Primary Sampling Units with CDL percent cultivation



Primary Sampling Units with percent crop(s) cover, overlaying a 2010 CDL image product

CDL based stratification of a NASS Area Sampling Frame (ASF)

- ASF stratification – computing the percent land covered by cultivation (all crops) within a PSU from CDL.
- ASF covariate stratification – computing the percent land covered by a specific crop within a PSU from CDL!
- Classifying each PSU into a defined stratum based on percent covered and the stratum definition.



Automated Covariate Stratification Procedure

1) Derive state level covariate data sets:
   - ❑ Select a region of interest from (2007-2010) CDL data, such as a state;
   - ❑ Combine the specific crop(s) (i.e. corn/soy, wheat or cotton) from CDL data over multi-year (2007-2010) 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";
   - ❑ Save the resulting data into a new covariate data layer ;
2) Load and overlay an individual ASF PSU boundary on the CDL covariate data layer;
3) 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.
4) Map each PSU into a defined stratum based on percent covariate covered and the stratum definition.
5) For better efficiency, sub-stratification should be conducted:

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

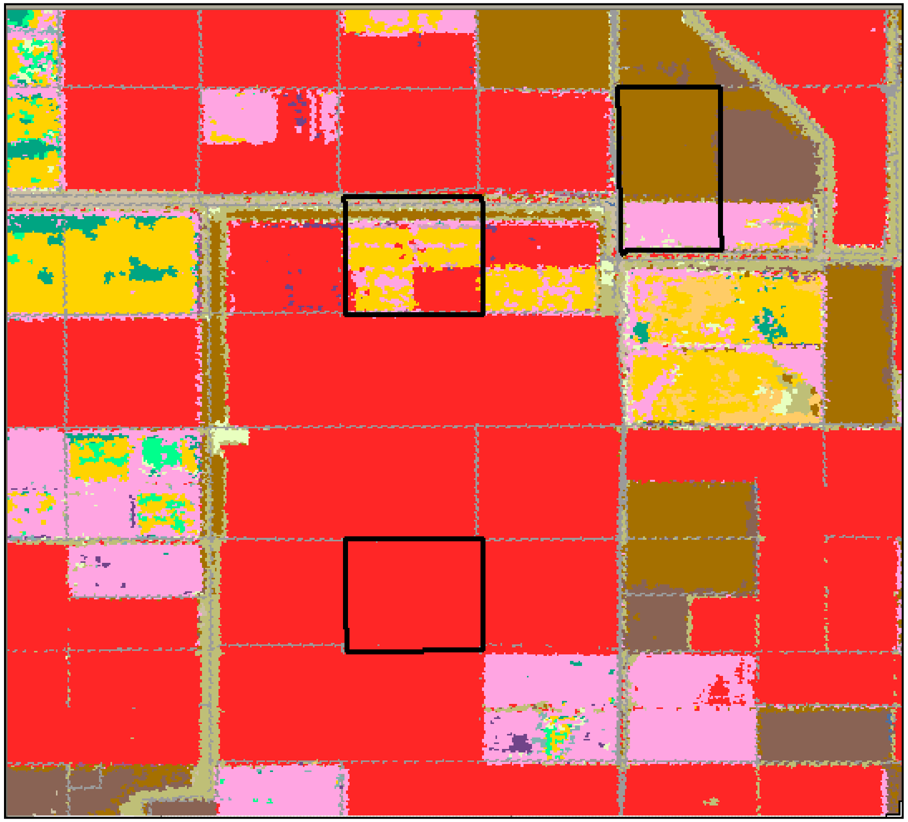
$$T_j \geq \sum_{h=1}^{H} \frac{N_h^2 S_{h,j}^2}{n_h}$$

- Where T_j is a target variance for commodity j for stratum h in {1,…,H}

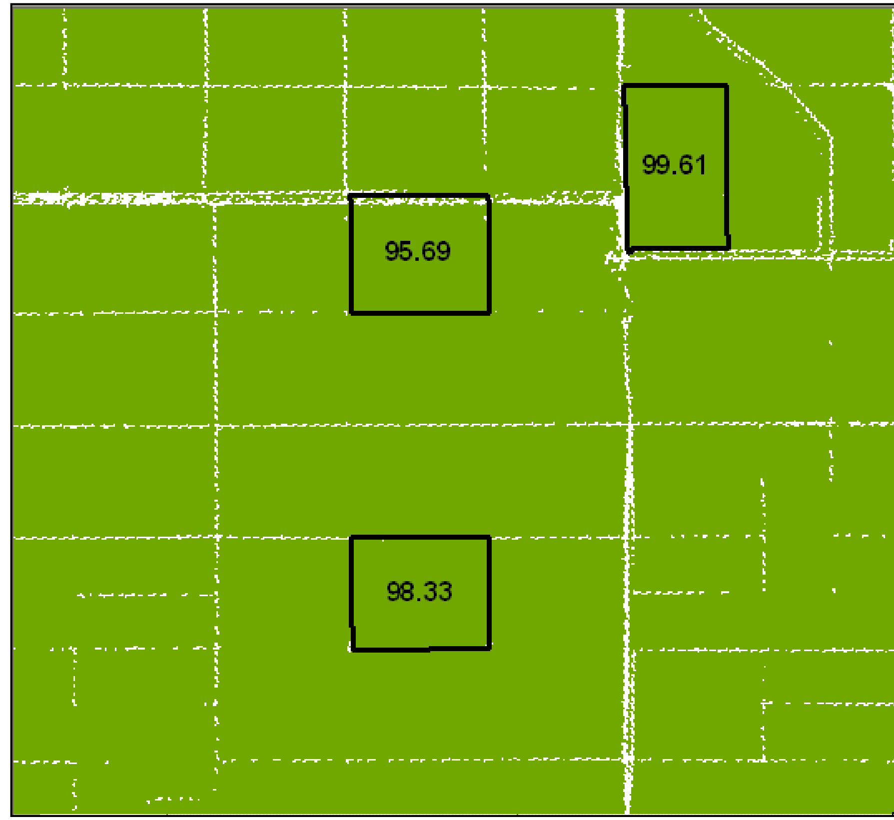
$$S_{h,j}^2 = (N_h - 1)^{-1} \sum_{i=1}^{N_h} (x_{i,j} - \bar{x}_j)_{h,j}^2$$

Stratification Results

CDL (2007-2010) covariates at the 2011 June Agricultural Survey (JAS) segment level

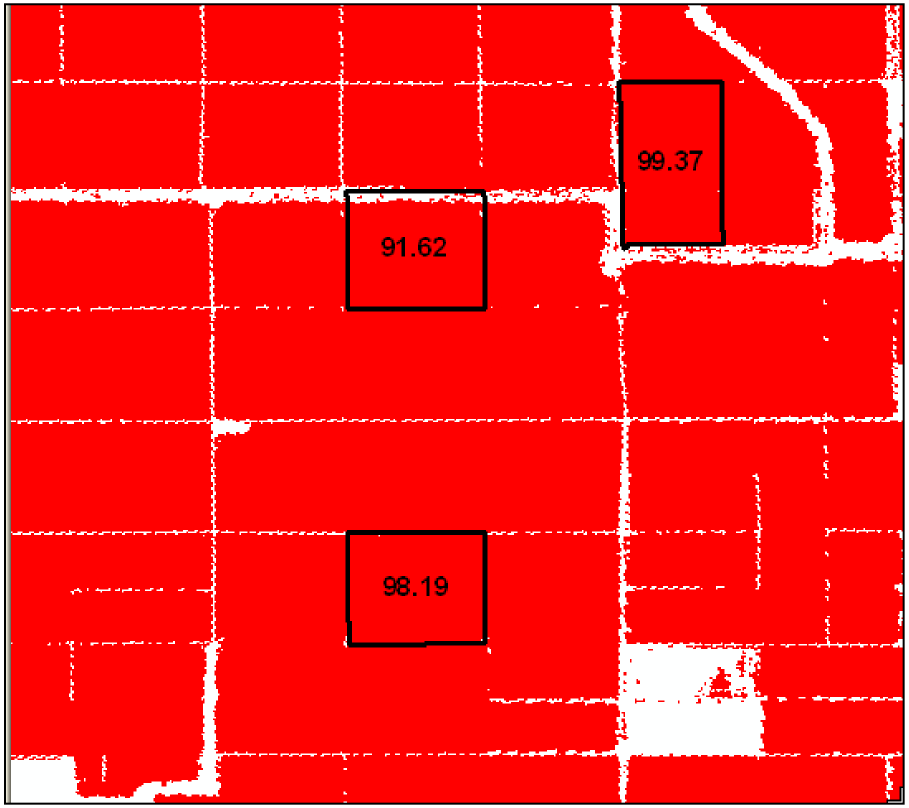


CA11 Cropland Data Layer with JAS segments

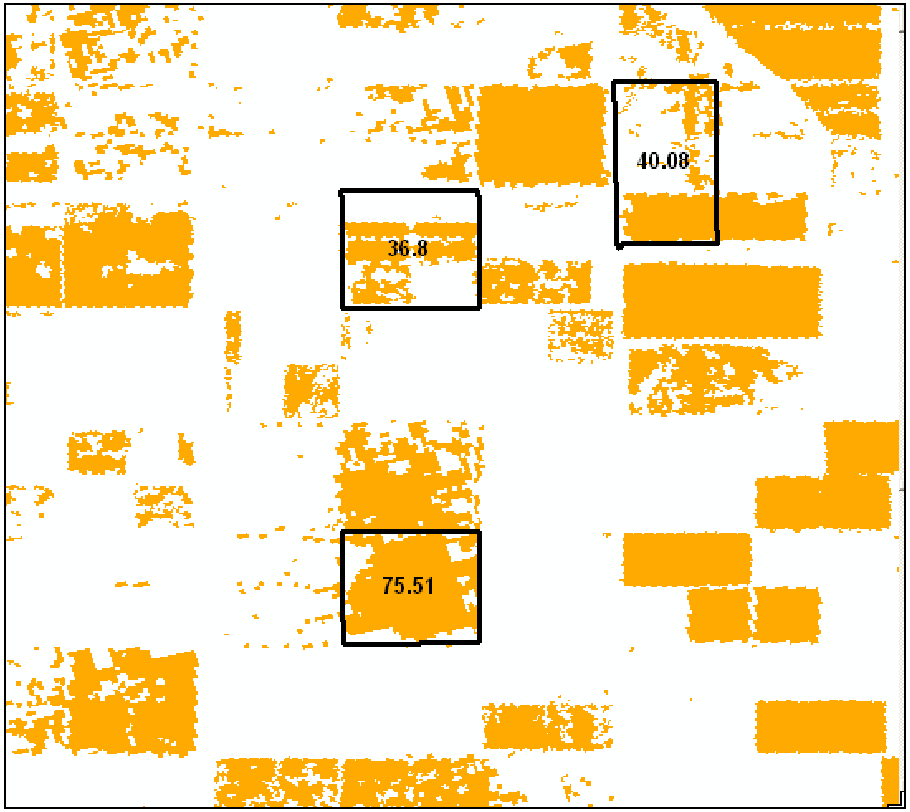


Multi Year (2007-2010) cultivated data set with JAS segments (**percent cultivation** calculated)

CDL (2007-2010) covariates at the 2011 June Agricultural Survey (JAS) segment level



Multi-year (2007-2010) **cotton** data set



Multi Year (2007-2010) **corn/soy** data set

CDL Covariate Predictive Accuracy

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

Validation : 2011 Farm Service Agency Common Land Unit/ NLCD 2006 – cultivation
2011 Cropland Data Layers - corn/soy, wheat, cotton

Ultimate Effect Assessment of Applications of Covariate Data

Direct Applications of CDL Covariate Data within NASS

- In the past, commodity information was derived from NASS county level statistics 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.

Ultimate Assessment - 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 defined by the** variance of the estimator from sampling plan (stratified covariate based sampling) divided by the variance of the 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

Design Effects

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

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

Conclusion

- The strength of the NASS Cropland Data Layer (CDL) product and CDL based stratification method is **objective and consistent in 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**.

Working in Progress

1)Derive covariate data sets from crop planting intensity at each pixel during a multi-year (2008-2012) CDL data, not simple crop coverage.

2) Compute percent covariate cover of each ASF PSU with intensity weighting for every pixel.

Thank you

Questions?

Claire.Boryan@nass.usda.gov
Zhengwei.Yang@nass.usda.gov
USDA/NASS/RDD
Feb 22, 2012

Procedure for Deriving *Intensity* Covariate Based on Cropland Data Layer (CDL)

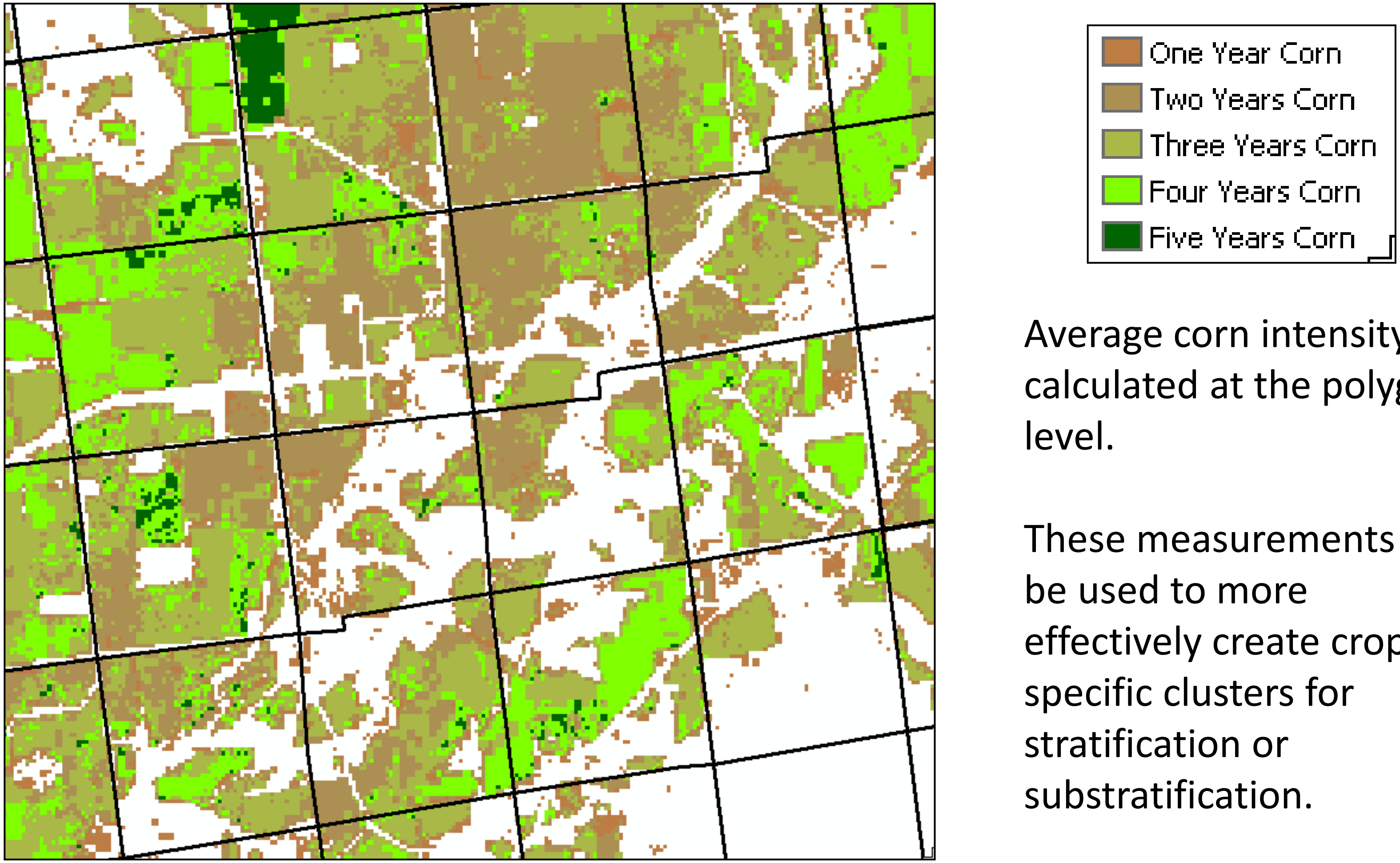
1) Derive state level covariate data sets from multi-year (2008-2012) CDL data by combining the specific crop (i.e. corn, soy, wheat or cotton) at the pixel level over five year period into five intensity categories and assigning the corresponding pixels with a value of "0,""1", "2", "3", "4" and "5" indicating the number of years a pixel was planted to the specific crop.

2) Load and overlay an individual ASF PSU boundary or grid on the CDL covariate layer;

3) Compute percent intensity covariate of each ASF PSU by summing the total number of pixels with values "1 - 5" and the total number of all pixels within the PSU or grid boundary. The percent intensity for a specific crop is given by sum of the pixels divided by total number of pixels.

Comment: Stratification process is similar to regular covariate stratification. But the stratum definition has to be different.

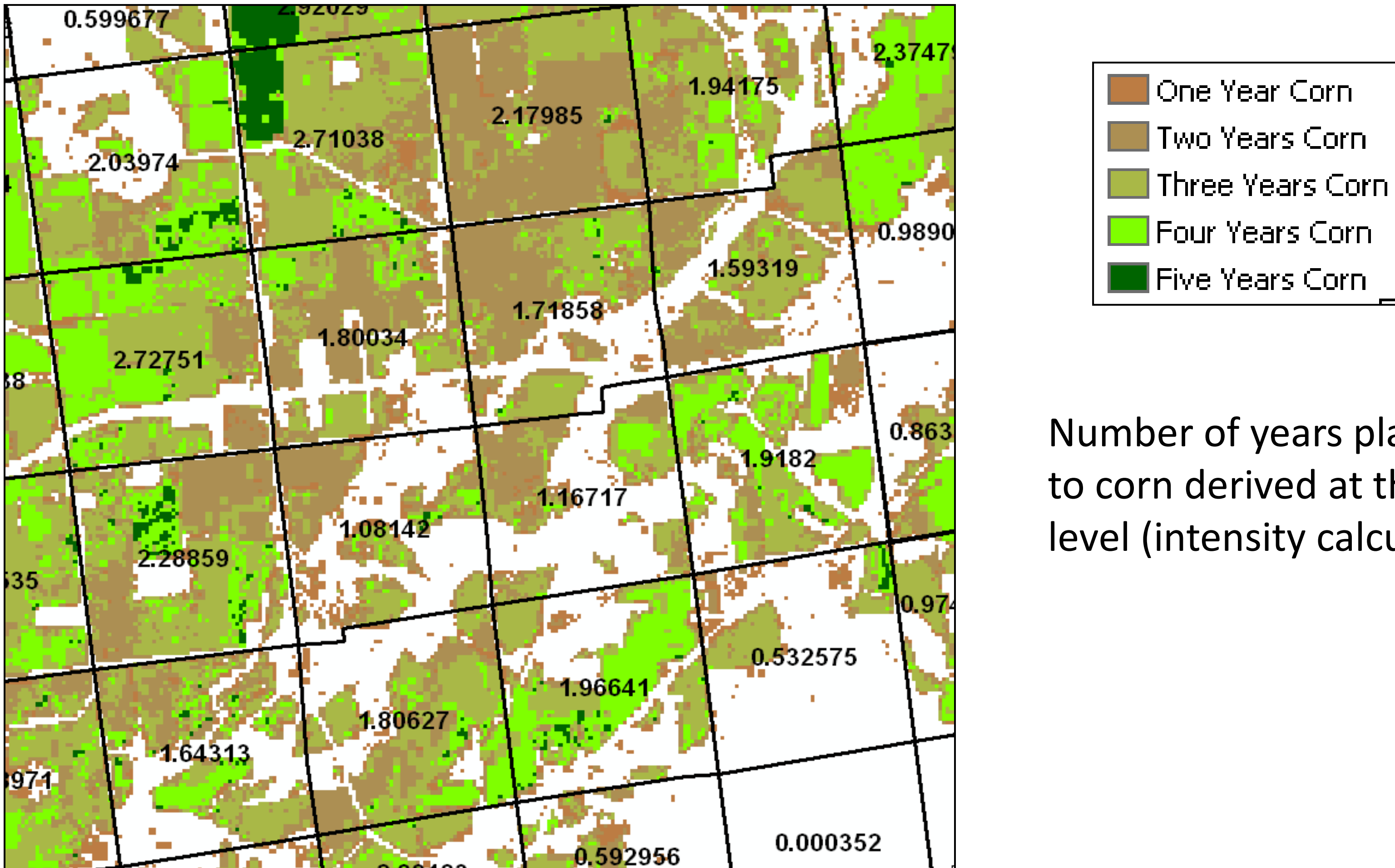
Indiana CDL (2008-2012) intensity covariates provide improved probability measure



Average corn intensity is calculated at the polygon level.

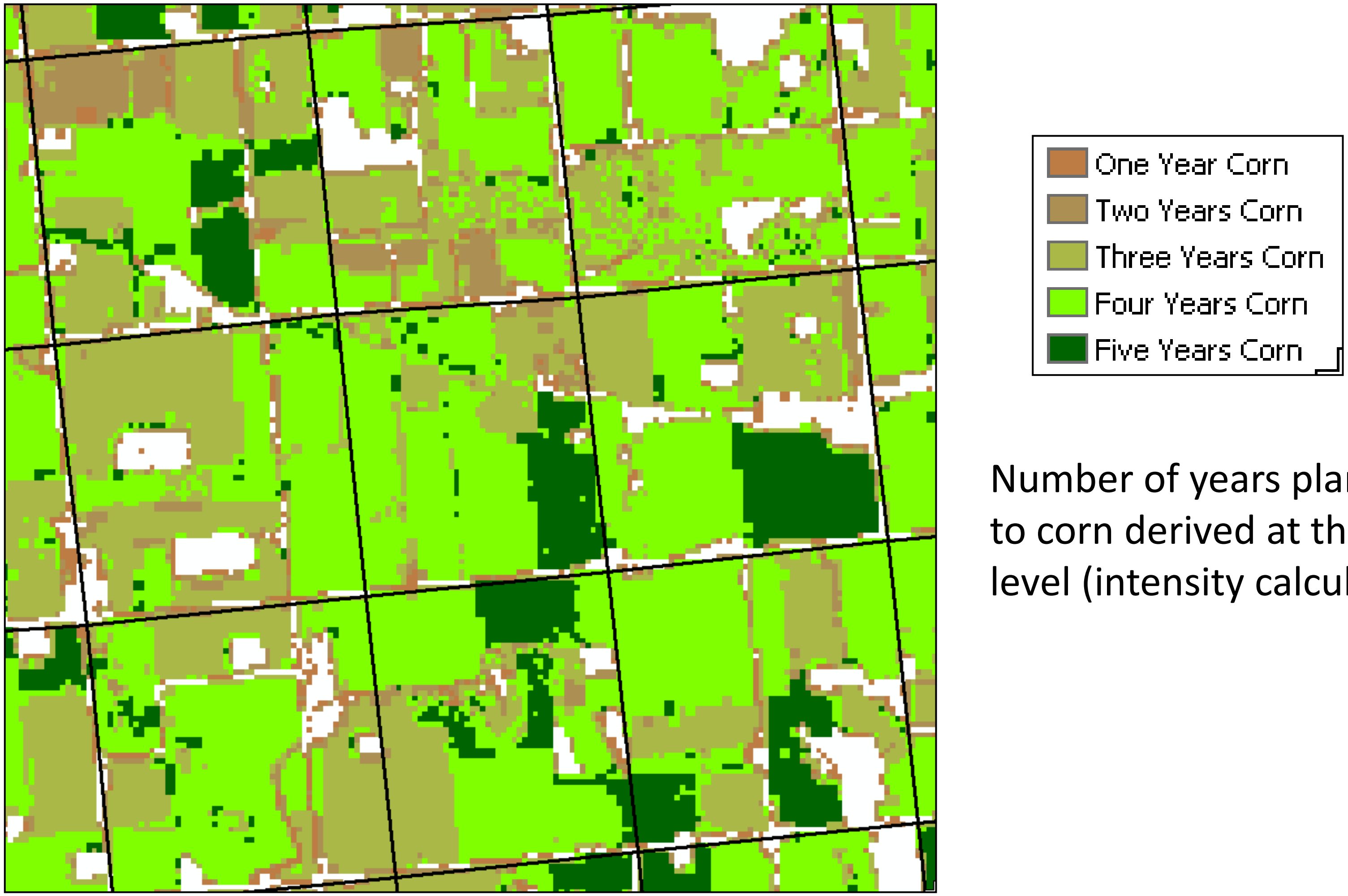
These measurements can be used to more effectively create crop specific clusters for stratification or substratification.

Indiana CDL (2008-2012) intensity covariates provide improved probability measure



Number of years planted to corn derived at the pixel level (intensity calculation)

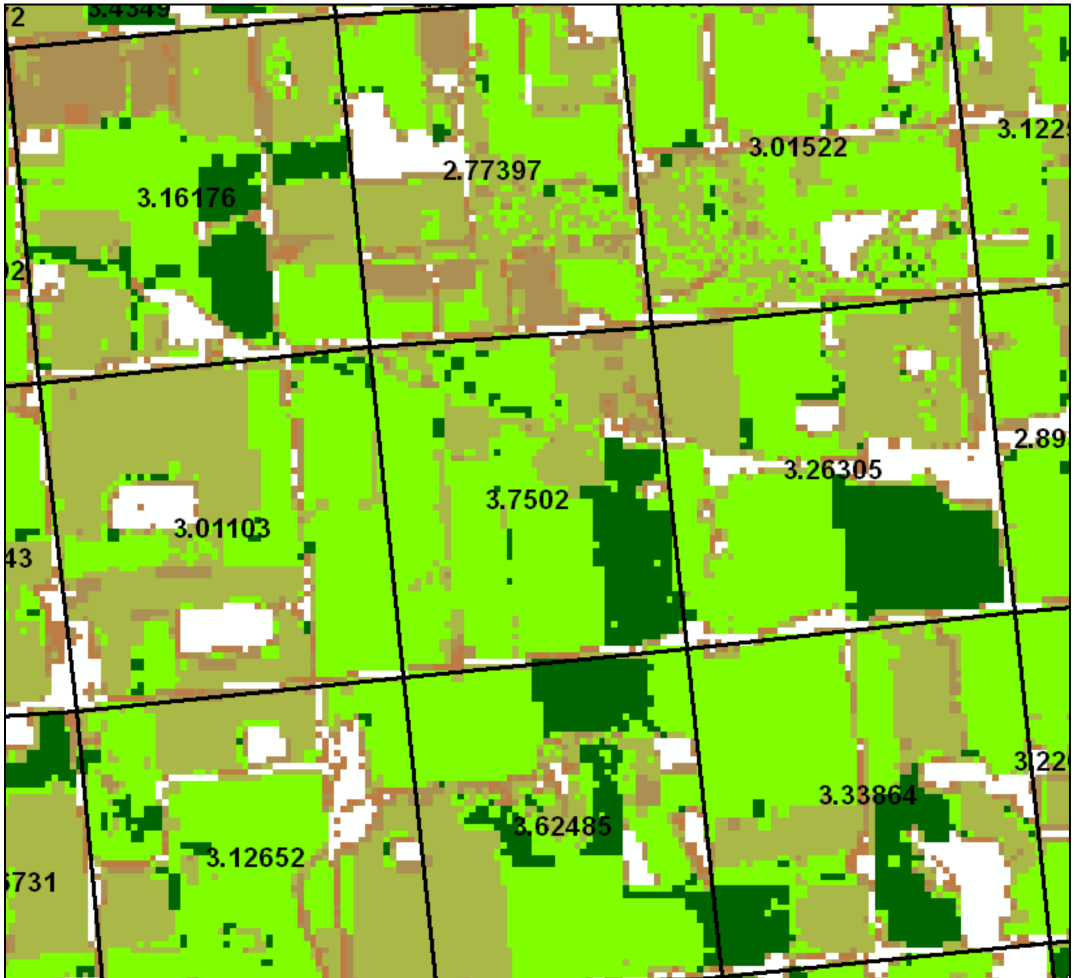
Indiana CDL (2008-2012) intensity covariates provide improved probability measure



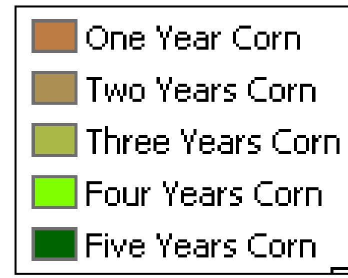
Area of increased corn planting intensity

Number of years planted to corn derived at the pixel level (intensity calculation)

Indiana CDL (2008-2012) intensity covariates provide improved probability measure



Area of increased corn planting intensity



Number of years planted to corn derived at the pixel level (intensity calculation)