Combining Estimates from Several Sources for Estimating Acreage of Crops

Stephanie Zimmer[1] Jae-Kwang Kim[1] Sarah Nusser[1] Cindy Yu[1]
Zhengyuan Zhu[1] Shu Yang[1] Michael Price[1] Jeff Bailey[2]
Jonathan Lisic[2]

[1]Iowa State University

[2]USDA National Agricultural Statistics Service

June 3, 2013

This research was partially supported by USDA NASS cooperative agreement 58-3AEU-1-0012.

June Area Survey

- National Agricultural Statistics Service (NASS) is committed to providing timely, accurate, and useful statistics in service to U.S. agriculture
- June Area Survey (JAS) uses an area frame to sample land to supply direct estimates of acreage and measures of sampling coverage for the nation and selected states
- JAS has small sample size which creates large sampling variance
- We propose using two other sources of data combined with JAS to reduce variance of crop acreage estimates

What We are Estimating

- Will focus on acreage estimates of crops today
- Estimate acreage of crops at county level
- Major crops include corn, soybeans, cotton, winter wheat, spring wheat, durum wheat which are all forms of cultivated crop land
- Goal is to give improved crop acreage estimate at state levels. In order to combine information, we need estimates in the county levels.

Combining Sources of Data

- Survey data, administrative data, and geospatial data source
- Goal is to generate unbiased acreage estimates from each source
- By combining them, we have a single unbiased estimate of crop acreage with more precision than any source alone
- Each have their own sources of errors, expressed via statistical models which will be discussed

JAS Data and Error Structure

- Data collected in first two weeks of June via personal interview from an area sample
- JAS acreage data have multiple sources of error
  - Nonresponse error - adjusted for in weights
  - Measurement error - not modeled in this work but believed to be relatively small and random
  - Sampling error - largest source of error

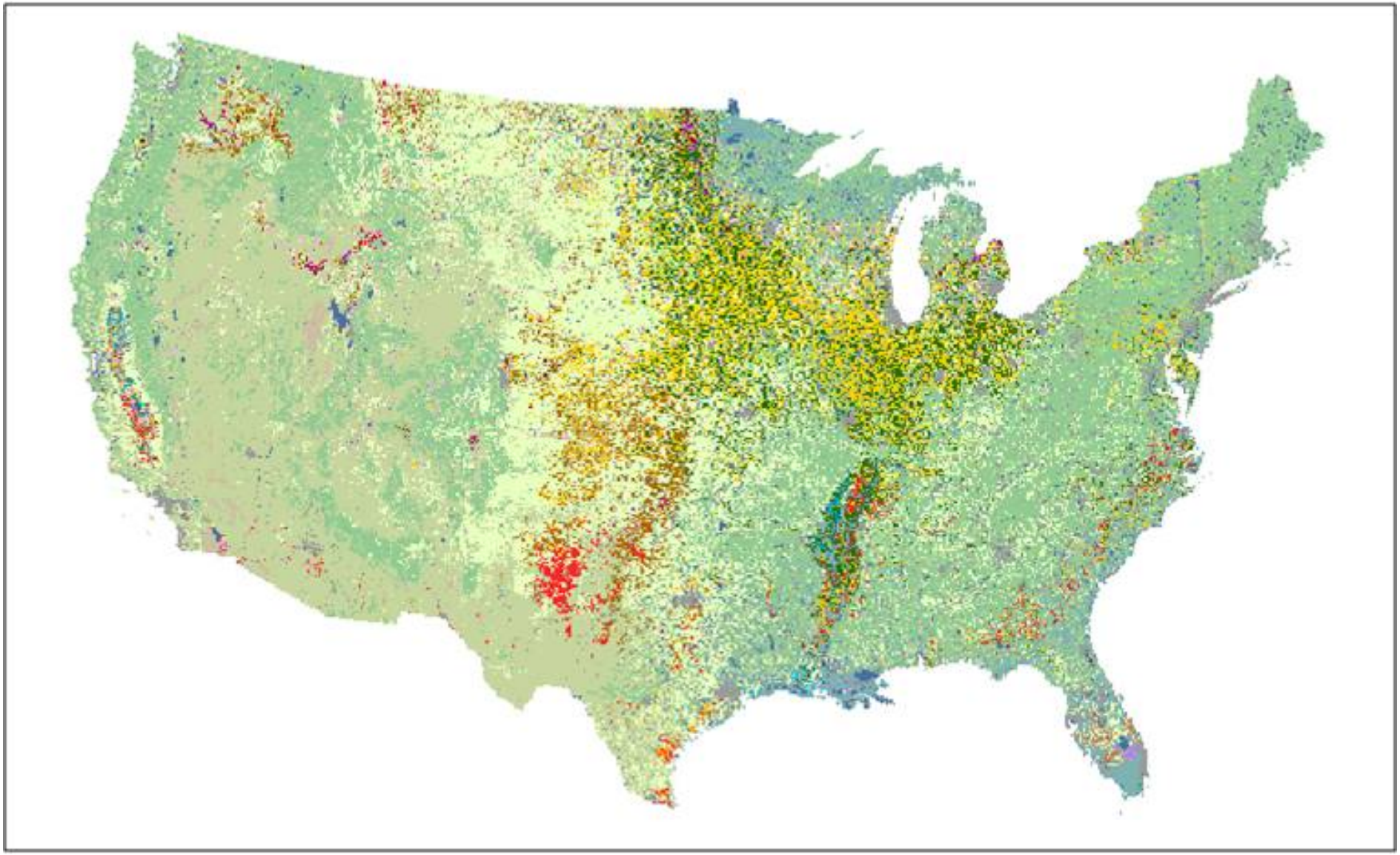
Farm Service Agency (FSA) - Another Source of Data

- FSA collects info from farmers on field boundaries and crops grown in fields
- Farmers register for farm programs or insurance certification
- Farmers register fields with FSA voluntarily so not all land is registered (undercoverage)
- If we can estimate undercoverage by using propensity scores, we can create an unbiased estimator of acreage
- Michael Price (Poster Session) is working on estimating and adjusting for this undercoverage
- Variance of adjusted acreage estimate is estimated using jackknife method

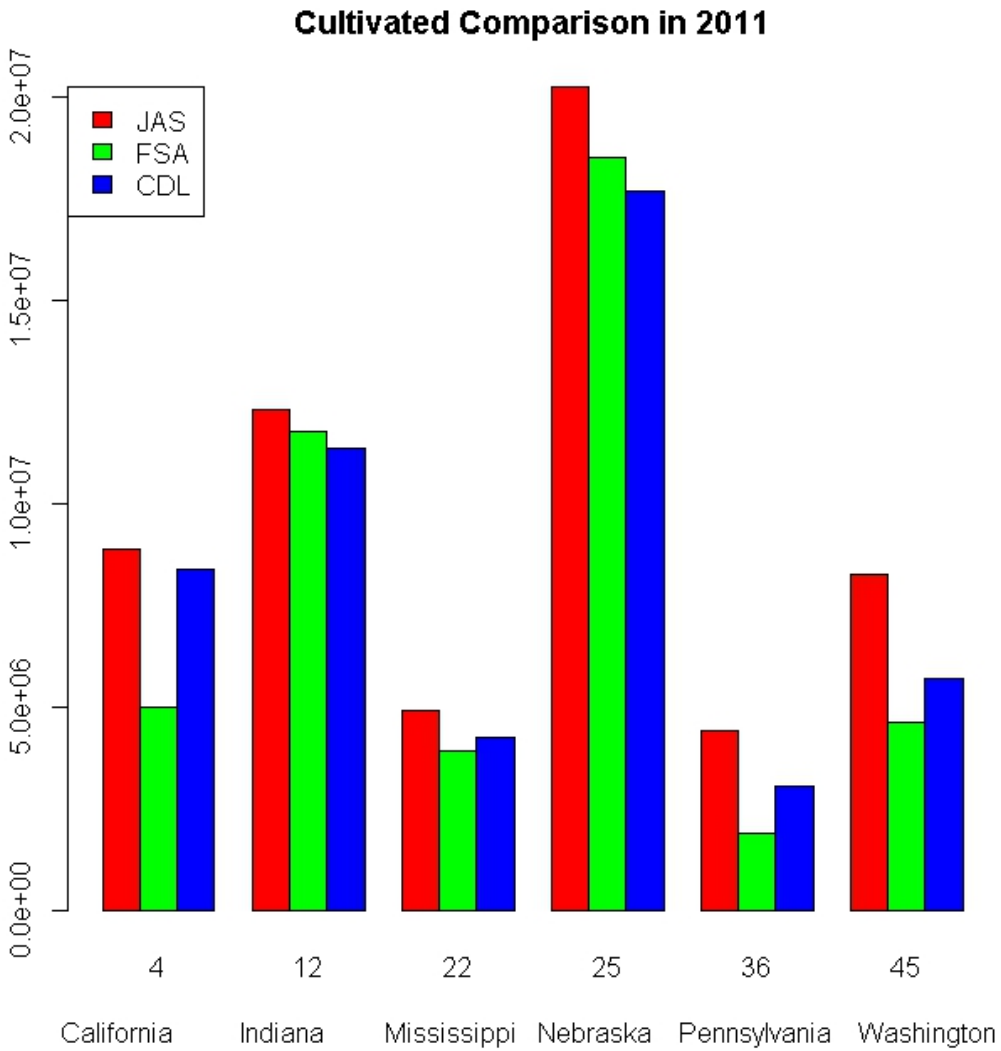
Cropland Data Layer (CDL) - Another Source of Data

- Classification of satellite imagery to reflect land cover; categories include corn, soybean, wheat, forest, urban, etc.
- Uses data from FSA and other sources as ground truth
- Full coverage, but subject to classification error
- Accuracy high at state level, but less accurate for smaller area
- Variance of CDL estimate is being modeled as part of project

CDL



Cultivated Cropland Comparison of Three Estimates



Basic Setup

- Crop acre estimates for end of growing season for each county i in each state
- X_i denotes the true acreage of a crop
- County level crop acre estimates
  - JAS estimate: $\hat{X}_i$ and $\hat{V}(\hat{X}_i)$
  - FSA estimate $\hat{Y}_{1i}$ and $\hat{V}(\hat{Y}_{1i})$
  - CDL estimate: $\hat{Y}_{2i}$ and $\hat{V}(\hat{Y}_{2i})$
- Since JAS sample is spatially sparse, not all counties have direct estimates so won't use these counties in parameter estimation but will use in prediction
- Our goal is to obtain a best prediction of X_i
- Estimate model parameters
- Generate county-level estimates (predictions) for each source

Model for JAS Data

JAS data is subject to sampling error so we use a sampling error model

$$\hat{X}_i = X_i + u_i$$

whereX_i is true acreage and $u_i \sim \left(0, \hat{V}(\hat{X}_i)\right)$, and

$$\hat{X}_i = \sum_{j \in S_i} w_{ij} X_{ij}$$

where w_{ij} is the nonresponse adjusted sampling weight X_{ij} is reported acreage of crop in sample j, county i
$\hat{V}(\hat{X}_i)$ is estimated using design-based estimator (2-stage stratified sample)

Models for FSA and CDL Data

- Each source has two models: structural and measurement error
- Structural error model is about the survey population
- Measurement error model is about the estimates
- We assume the measurement variances (sampling variances) are estimated relatively accurately - use smoothing techniques to reduce variability in variance estimates
- We use parametric models for the structural error model.

Model for FSA Data

1. Structural error model:

$$Y_{1i} = \beta_0 + \beta_1 X_i + e_{1i}$$

   where $e_{1i} \sim (0, \sigma_1^2)$

2. Measurement error model: $\hat{Y}_{1i} = Y_{1i} + u_{1i}$ where $u_{1i} \sim \left(0, \hat{V}(\hat{Y}_{1i})\right)$.

$\hat{Y}_{1i}$ is the undercoverage-adjusted crop acreage estimate and $\hat{V}(\hat{Y}_{1i})$ is the estimated variance for county i

σ_1^2 represents the lack of fit when explaining population values of FSA (Y_{1i}) by population values of JAS (X_i)

Model for CDL Data

1. For example, simple linear regression model: Structural error model:

$$Y_{2i} = \beta_0^* + \beta_1^* X_i + e_{2i}$$

where $e_{2i} \sim (0, \sigma_2^2)$

2. Measurement error model:

$$\hat{Y}_{2i} = Y_{2i} + u_{2i}$$

where $u_{2i} \sim \left(0, \hat{V}(\hat{Y}_{2i})\right)$.

$\hat{Y}_{2i} = \sum_{j \in A_i} \delta_{ij} a_{ij}$ where δ_{ij} is the indicator of the crop in pixel j in county i, A_i is the set of pixels of county i and a_{ij} is the area of the pixel, $\hat{V}(\hat{Y}_{2i})$ is being estimated in a separate project

σ_2^2 represents the lack of fit when explaining population values of CDL by population values of JAS

Combined Model

$$\begin{pmatrix} \hat{Y}_{1i} - \beta_0 \\ \hat{Y}_{2i} - \beta_0^* \\ \hat{X}_i \end{pmatrix} = \begin{pmatrix} \beta_1 \\ \beta_1^* \\ 1 \end{pmatrix} X_i + \begin{pmatrix} e_{1i} + u_{1i} \\ e_{2i} + u_{2i} \\ u_i \end{pmatrix}$$

where

$$\begin{pmatrix} e_{1i} + u_{1i} \\ e_{2i} + u_{2i} \\ u_i \end{pmatrix} \sim \left[\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \hat{V}(\hat{Y}_{1i}) + \sigma_1^2 & 0 & 0 \\ 0 & \hat{V}(\hat{Y}_{2i}) + \sigma_2^2 & 0 \\ 0 & 0 & \hat{V}(\hat{X}_i) \end{pmatrix} \right]$$

$(\beta_0, \beta_1, \sigma_1^2, \beta_0^*, \beta_1^*, \sigma_2^2)$ can be estimated via Method of Moments or MLE, we propose using PFI to estimate MLEs of parameters (details omitted)

After Parameter Estimation: Prediction Using GLS

$$\begin{pmatrix} \hat{Y}_{1i} - \hat{\beta}_0 \\ \hat{Y}_{2i} - \hat{\beta}_0^* \\ \hat{X}_i \end{pmatrix} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_1^* \\ 1 \end{pmatrix} X_i + \begin{pmatrix} e_{1i} + u_{1i} \\ e_{2i} + u_{2i} \\ u_i \end{pmatrix}$$

where

$$\begin{pmatrix} e_{1i} + u_{1i} \\ e_{2i} + u_{2i} \\ u_i \end{pmatrix} \sim \left[\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \hat{V}(\hat{Y}_{1i}) + \hat{\sigma}_1^2 & 0 & 0 \\ 0 & \hat{V}(\hat{Y}_{2i}) + \hat{\sigma}_2^2 & 0 \\ 0 & 0 & \hat{V}(\hat{X}_i) \end{pmatrix} \right]$$

Thus we can express

$$Y = Z\theta + e$$

where $e \sim (0, \hat{V})$ and obtain the best prediction of X_i as

$$\hat{\theta} = \left(Z' \hat{V}^{-1} Z \right)^{-1} Z' \hat{V}^{-1} Y$$

Comments on assumptions

- The assumption of $\mathrm{Cov}(e_{1i}, e_{2i}) = 0$ is not true but still get consistent estimator of $\hat{\theta}$
- In some counties, we may not have $\hat{X}_i$ due to small sample size and will simplify estimation to

$$\begin{pmatrix} \hat{Y}_{1i} - \hat{\beta}_0 \\ \hat{Y}_{2i} - \hat{\beta}_0^* \end{pmatrix} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_1^* \end{pmatrix} X_i + \begin{pmatrix} e_{1i} + u_{1i} \\ e_{2i} + u_{2i} \end{pmatrix}$$

where we apply GLS estimator to this

Conclusion

- Area level estimates (county level estimates) are computed from several source in order to combine information
- Structural error model is important to link the different data sources - describes sampling error for JAS and error in data sources for CDL/FSA
- We only observe sample estimates so measurement error model is also needed for combining sources
- We didn't, but could, incorporate other sources of error into each source of data

Future Work and Discussion

- Is assumption of linear models adequate? We have 0 inflated data
- Smaller crops will be a challenge as $\hat{X}_i$ will be 0 frequently
- We would like to monthly update estimates, is this feasible? Models may change, error structure may change. Interim data is of poorer quality for FSA and CDL
- How do administrative data sources fit into TSE paradigm?