Using administrative data to produce official statistics: an application to end-of-season acreage estimation

Andreea L. Erciulescu

National Institute of Statistical Sciences
USDA National Agricultural Statistics Service

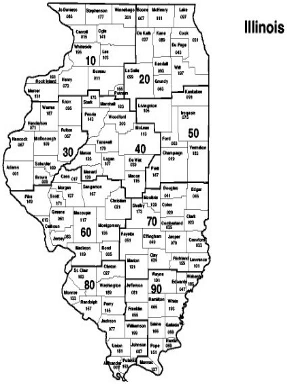
Symposium on Data Science and Statistics
May 17, 2018

Acknowledgements

Nathan Cruze
Habtamu Benecha
Valbona Bejleri
Balgobin Nandram
Claire Boryan
Rick Mueller
Wendy Barboza
Linda Young
Nell Sedransk

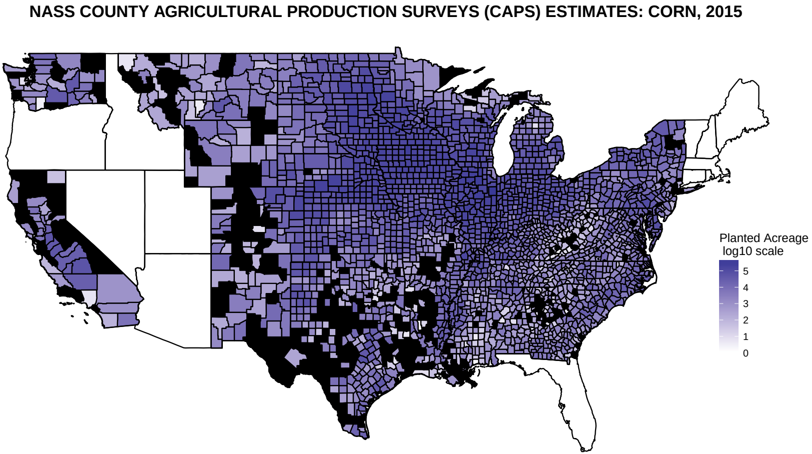
Motivation: End-of-Season Estimates

- USDA National Agricultural Statistics Service (NASS)
  - 400+ reports annually, including crops estimates
- Agricultural Statistics Board
  - Expert Assessment
  - State
    - Agricultural Statistics District (ASD), County
  - Publication standard
    - 30+ positive reports for yield *or*
    - 3+ positive reports for yield and 25%+ coverage for harvested acreage
  - NASS QuickStats
- Two major users within USDA
  - Farm Service Agency (FSA)
  - Risk Management Agency (RMA)



NASS county estimates are used in the process of setting payments for some agricultural programs!

Motivation: County-Level Planted Acreage Estimates



- 2837 counties in 36 sampled states
- 2426 in-sample counties and 411 not-in-sample counties

Motivation and Goals

- Explore auxiliary sources that indicate corn planting activity
  - list-based survey; changes in planting practices
  - each survey response includes information on entire farm or ranch, all commodities
  - approach: commodity-specific administrative data sources
- Combine survey and auxiliary data to produce substate-level* predictions and measures of uncertainty for in-sample and not-in-sample domains
  - small sample sizes (number of positive reports used to produce the survey summary)
  - approach: small area models
- Preserve agreement between different aggregation levels

*county-level and (agricultural statistics) district-level

Using Information from Multiple Data Sources

Table 1: Counties, *in Sampled States*, with Corn Planting Activity, 2015

Data Source (USDA)	**Data Collection Method**	**Number of Counties**
NASS CAPS	Probability Sample	2426
Farm Service Agency (FSA)	Volunteer Reporting	2398
Risk Management Agency (RMA)	Volunteer Reporting	2230
NASS Cropland Data Layer (CDL)	Remote Sensing + Ground-Reference	2761

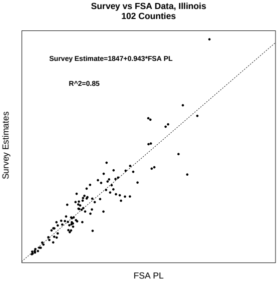
- Define Set of Counties with Corn Planting Activity
  - combine NASS CAPS, FSA, RMA and CDL

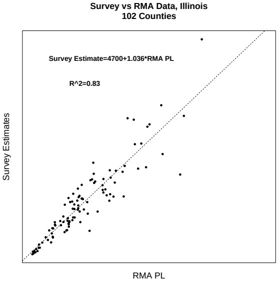
Small Amount of Survey Summary Data
2015 Corn Planted Acreage

Nationwide summaries

- sample size within a county: $[1, 191]$; median 18
- sample size within a district: $[1, 993]$; median 206
- number of districts within a state: $[3, 15]$; median 9
- number of counties within a district: $[1, 32]$; median 8

Exploring Relationships between Multiple Data Sources 2015 Corn Planted Acreage (PL); County-Level





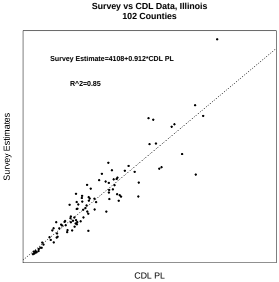


Table 2: Nationwide Summaries

	FSA PL			RMA PL			CDL PL		
	1st Qu.	Median	3rd Qu.	1st Qu.	Median	3rd Qu.	1st Qu.	Median	3rd Qu.
R^2	0.82	0.89	0.92	0.76	0.86	0.91	0.85	0.90	0.93

Borrowing Information from Multiple Data Sources 2015 Corn Planted Acreage (PL); County-Level

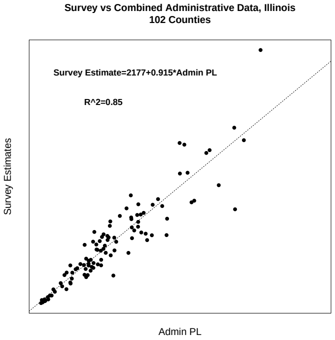


Table 3: Nationwide Summaries

	FSA PL			RMA PL			CDL PL			**Admin PL**		
	1st Qu.	Median	3rd Qu.	1st Qu.	Median	3rd Qu.	1st Qu.	Median	3rd Qu.	1st Qu.	Median	3rd Qu.
R^2	0.82	0.89	0.92	0.76	0.86	0.91	0.85	0.90	0.93	0.85	0.90	0.93

Admin PL: combine FSA, RMA and CDL, with preference for maximum planted acreage

Approach: Subarea-Level Model for a Given State

Linkage model

$$\begin{aligned} \theta_{ij}|(\boldsymbol{\beta}, \sigma_u^2, v_i) &\sim N(\mathbf{x}_{ij}'\boldsymbol{\beta} + v_i, \sigma_u^2) \\ v_i|\sigma_v^2 &\sim N(0, \sigma_v^2) \end{aligned}$$

Sampling model

$$\hat{\theta}_{ij}|(\theta_{ij}, \hat{\sigma}_{ij}^2) \sim N(\theta_{ij}, \hat{\sigma}_{ij}^2)$$

Prior distributions

$$\pi(\boldsymbol{\beta}, \sigma_u^2, \sigma_v^2) = \pi(\boldsymbol{\beta})\pi(\sigma_u^2)\pi(\sigma_v^2)$$

- $i = 1, ..., m$, areas (districts)
- $j = 1, ..., n_i^c$, subareas (counties) in area (district) i
- $\sum_{i=1}^{m} n_i^c = n^c$, number of counties
- θ_{ij}, county-level parameter of interest
- $(\hat{\theta}_{ij}, \hat{\sigma}_{ij}^2)$, survey summary
- $\mathbf{x}_{ij} = (1, x_{ij})$
- x_{ij} = Admin PL (M); for comparison, NULL (M0) and Admin PL as combined FSA and RMA only (M1) are also used

Modeling Strategies with Incomplete Data

Missing x_{ij}, but available $\hat{\theta}_{ij}$

- impute x_{ij} using the administrative data available for a similar county in the given state
  - absolute-value norm, applied to the corresponding $\hat{\theta}_{ij}$'s

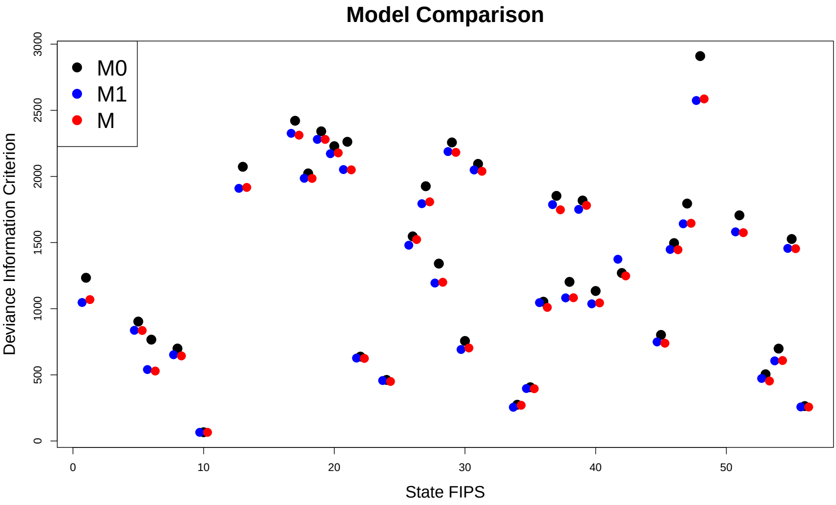
Available $(\hat{\theta}_{ij}, \hat{\sigma}^2_{ij}, x_{ij})$

- posterior summaries using MCMC iterates (after burn-in and thinning); $r = 1, ..., R$
  - parameter iterates: $\boldsymbol{\beta}_r, \sigma^2_{u,r}, \sigma^2_{v,r}$
  - county-level iterates: $\theta_{ij,r}$
  - district-level iterates: $\theta_{i,r} := \sum_{j=1}^{n_i^c} \theta_{ij,r}$

Missing $(\hat{\theta}_{ij}, \hat{\sigma}^2_{ij})$, but x_{ij} available

- prediction using the linkage model: $\theta_{ij,r} \sim N(\mathbf{x}_{ij}^{'}\boldsymbol{\beta}_r + v_{i,r}, \sigma^2_{u,r})$

Results: Model Comparison



Results: Shrinkage away from the Survey Estimate

Posterior mean:

$$\begin{aligned}\tilde{\theta}_{ij} &= \mathbf{x}_{ij}^{'}\tilde{\boldsymbol{\beta}} + \tilde{\gamma}_i(\bar{\hat{\theta}}_i^{\gamma} - \bar{\mathbf{x}}_i^{\gamma'}\tilde{\boldsymbol{\beta}}) + \tilde{\gamma}_{ij}\left\{\hat{\theta}_{ij} - \mathbf{x}_{ij}^{'}\tilde{\boldsymbol{\beta}} - \tilde{\gamma}_i(\bar{\hat{\theta}}_i^{\gamma} - \bar{\mathbf{x}}_i^{\gamma'}\tilde{\boldsymbol{\beta}})\right\} \\ &= \tilde{\gamma}_{ij}\hat{\theta}_{ij} + (1-\tilde{\gamma}_{ij})\left\{\mathbf{x}_{ij}^{'}\tilde{\boldsymbol{\beta}} + \tilde{\gamma}_i(\bar{\hat{\theta}}_i^{\gamma} - \bar{\mathbf{x}}_i^{\gamma'}\tilde{\boldsymbol{\beta}})\right\}\end{aligned}$$

- $\tilde{\gamma}_{ij} = \frac{\tilde{\sigma}_u^2}{\tilde{\sigma}_u^2 + \hat{\sigma}_{ij}^2}$, $\tilde{\gamma}_{i.} = \sum_{j=1}^{n_i^c}\tilde{\gamma}_{ij}$, $\tilde{\gamma}_i = \frac{\tilde{\sigma}_v^2}{\tilde{\sigma}_v^2 + \tilde{\sigma}_u^2(\tilde{\gamma}_{i.})^{-1}}$
- $\bar{\hat{\theta}}_i^{\gamma} = (\tilde{\gamma}_{i.})^{-1}\sum_{j=1}^{n_i^c}\tilde{\gamma}_{ij}\hat{\theta}_{ij}$, $\bar{x}_i^{\gamma} = (\tilde{\gamma}_{i.})^{-1}\sum_{j=1}^{n_i^c}\tilde{\gamma}_{ij}x_{ij}$

Table 4: Summary of Estimated Shrinkage Coefficients γ_{ij} (%)

Approach	Covariate ADMIN PL	1st Qu.	Median	3rd Qu.
Model M0	None	60.66	85.69	98.01
Model M1	FSA and RMA	2.67	11.41	44.92
Model M	**FSA, RMA and CDL**	2.42	10.25	40.94

Benchmarking Constraint

For a prepublished state-level value, a

- $\sum_{i,j}^{n^{c*}} \tilde{\theta}_{ij}^{B} = a$, n^{c*} is the total number of counties
- ratio adjustment, applied at the (MCMC) iteration-level

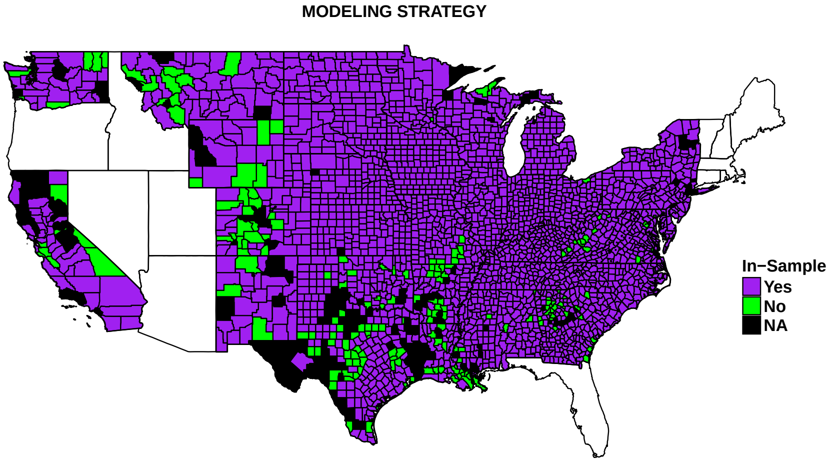
$$\theta_{ij,r}^{B} := \theta_{ij,r} \times a \times \left(\sum_{k=1}^{m} \sum_{l=1}^{n_k^{c*}} \theta_{kl,r} \right)^{-1},$$

n_k^{c*} is the total number of counties in district k, $k = 1, ..., m$.

Discussion:

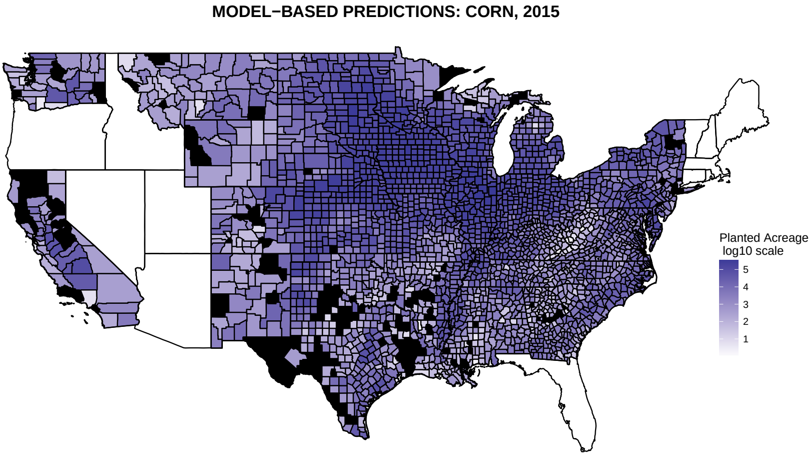
- defining the set of counties n^{c*}

Results



- 2420 in-sample counties and 209 not-in-sample counties (M)
    - Texas: largest number of not-in-sample predictions, 42 out of 184 counties, accounting for ∼0.7% of planted acreage in the state

Results: Increased Number of County-Level Estimates



- (M) model-based predictions available for 2629 counties
- RECALL: survey estimates available for 2426 counties

Results: Increased Precision

Table 5: SE Summaries for Counties with Available Survey Estimates

Approach	Covariate ADMIN PL	1st Qu.	Median	3rd Qu.
Survey		640.90	2719.00	9494.00
Model M1	FSA, RMA	429.40	1233.00	2850.00
Model M	**FSA, RMA and CDL**	429.30	1166.00	2839.00

Results: Decreased Relative Variability

Table 6: CV(%) Summaries for Counties with Available Survey Estimates

Approach	Covariate ADMIN PL	1st Qu.	Median	3rd Qu.
Survey		21.08	31.91	55.42
Model M1	FSA, RMA	5.97	12.60	38.74
Model M	**FSA, RMA and CDL**	5.90	11.84	37.92

Results: Official Statistics

- Composite predictions
- Common publication standard
  - 2420 counties with available survey estimates:
    - 1125 survey CVs $\leq$ 30% vs. 1693 model (M) CVs $\leq$ 30%
  - 2629 counties with available model-based (M) predictions:
    - 1696 model (M) CVs $\leq$ 30%
- Current NASS publication standard
  - county-level sample size and efficiency of weighting adjustments
  - 1622 counties published in NASS QuickStats

Summary and Future Work

Contributions of administrative data

- model-based county-level and district-level predictions, incorporating survey and administrative data (implicit weights)
- defined set of counties with planting activity
- reduction in the need for covariate imputation, by using remote sensing data (110(M1) vs. 12(M))
- increased number of county-level estimates (2486(M1) vs. 2629(M))
- increased precision and relative precision; model vs. survey
  - 2.67-71.39% / 19.96-74.5% in most of the county-level SE / CV
  - 18.27-58.59% / 28.72-62.55% in most of the district-level SE / CV
- official statistics

Future work

- out-of-sample states
- model specification; normality assumption and constraints
- quality of different data sources; imputation strategies and errors
- publication standard

References

Bell J., and Barboza W. (2012), "Evaluation of Using CVs as a Publication Standard." Paper presented at the Fourth International Conference on Establishment Surveys, Montreal, Quebec, Canada, June 11-14.

Cruze N.B., Erciulescu A.L., Nandram B., Barboza W.J., Young L.J. (2016), "Developments in Model-Based Estimation of County-Level Agricultural Estimates." *ICES V Proceedings. Alexandria, VA: American Statistical Association*.

Erciulescu A.L., Cruze N.B., Benecha H., Bejleri V., Nandram B. (2018), "On Increasing the Number of County-Level Crop Estimates." *FCSM Proceedings*. To appear.

Erciulescu A.L., Cruze N.B., Nandram B. (2016), "Model-Based County-Level Crop Estimates Incorporating Auxiliary Sources of Information." *JSM Proceedings. Survey Research Methods Section. Alexandria, VA: American Statistical Association*, 3591-3605.

Fay R.E. and Herriot R.A. (1979), "Estimates of income for small places: an application of James-Stein procedures to census data," *Journal of the American Statistical Association*, 74, 269-277.

Fuller W.A. and Goyeneche J.J. (1998), "Estimation of the state variance component," *Unpublished manuscript*.

Marker D. (2016), "Presentation to National Academy of Sciences Panel on Crop Estimates," *Unpublished presentation*. National Academy of Sciences report available at *https://www.nap.edu/catalog/24892/improving-crop-estimates-by-integrating-multiple-data-sources*.

Rao J.N.K. and Molina I. (2015), "Small Area Estimation," *Wiley Series in Survey Methodology*.

Torabi M. and Rao J.N.K. (2014), "On small area estimation under a sub-area level model," *Journal of Multivariate Analysis*, 127, 36-55.

USDA FSA (2014), "Farm Bill Home," *http://www.fsa.usda.gov/programs-and-services/farm-bill/index*.

USDA NASS (2016a), "Publications: Agricultural Statistics, Annual," *https://www.nass.usda.gov/Publications/Ag_Statistics*.

USDA NASS (2016b), "CropScape and Cropland Data Layer," *https://www.nass.usda.gov/Research_and_Science/Cropland/SARS1a.php*.

USDA NASS (2016c), "QuickStats," *https://quickstats.nass.usda.gov/*.

USDA NASS (2017a), "Crop Production Annual Summary," *http://usda.mannlib.cornell.edu/MannUsda/viewDocumentInfo.do?documentID=1047*.

USDA NASS (2017b), "Historical Track Record - Crop Production," *http://usda.mannlib.cornell.edu/MannUsda/viewDocumentInfo.do?documentID=1593*

USDA RMA (2014), "THE FARM BILL," *http://www.rma.usda.gov/news/currentissues/farmbill/*.

Thank you!

aerciulescu@niss.org

Internal Model Validation
Posterior Predictive Checks

- Posterior samples: $(\boldsymbol{\beta}^r, (\sigma_v^2)^r, (\sigma_u^2)^r), r = 1, ..., R$
- Draw replicates $(\theta_{ij}^t, y_{ij}^t), t = 1, ..., T$ (every 10^{th} sample from the R iterates):

$$\begin{aligned} v_i^t &\sim N(0, (\sigma_v^2)^t) \\ \theta_{ij}^t &\sim N(\mathbf{x}_{ij}^{'}\boldsymbol{\beta}^t + v_i^t, (\sigma_u^2)^t) \\ y_{ij}^t &\sim N(\theta_{ij}^t, (\hat{\sigma}_{ij}^2)^t) \end{aligned}$$

- For a given test statistic, i.e. identity function,

$$p = T^{-1} \sum_{t=1}^{T} I\left(T(y_{ij}^t) > T(\hat{\theta}_{ij})\right)$$

External Model Validation
NASS Official Values

- Agricultural Statistics Board and Census of Agriculture
- Five years: 2012-2016
- Multiple commodities: corn, soybeans, sorghum, wheat
- Comparison metrics: (absolute) (relative) differences, credible intervals coverage