On Increasing the Number of County-Level Crop Estimates

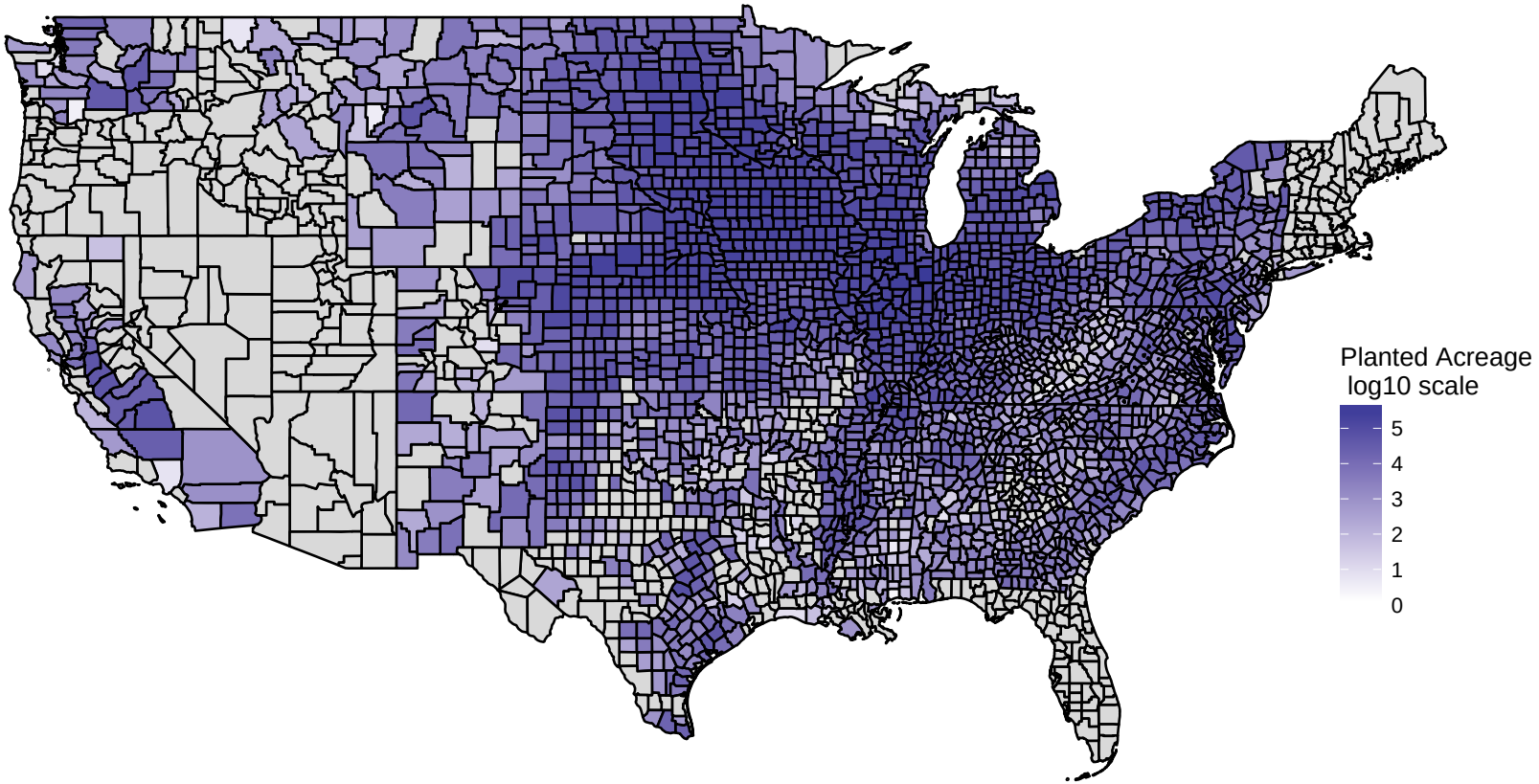
Andreea L. Erciulescu[1,2], Nathan Cruze[2], Habtamu Benecha[2], Valbona Bejleri[2], Balgobin Nandram[2,3]

1. National Institute of Statistical Sciences
2. USDA National Agricultural Statistics Service
3. Worcester Polytechnic Institute

Federal Committee on Statistical Methodology
Research and Policy Conference
March 8, 2018

Motivation: County-Level Planted Acreage Estimates

NASS COUNTY AGRICULTURAL PRODUCTION SURVEYS (CAPS) ESTIMATES: CORN, 2015



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

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

Motivation: Questions

- Are there ancillary sources that indicate corn planting activity in the 411 not-in-sample counties?
  - list-based survey; changes in planting practices
  - each survey response includes information on entire farm or ranch, all commodities
  - our approach: explore commodity-specific administrative data sources

- How to combine survey and auxiliary data to produce substate-level* estimates 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)
  - our approach: small area models

*county-level and (agricultural statistics) district-level, where a district is represented by a set of neighboring counties within a state

Using Information from Multiple Data Sources

Table 1: Number of Counties 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	2232

- Define Set of Counties with Corn Planting Activity
  - combine NASS CAPS, FSA and RMA $\rightarrow$ 2510 counties

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

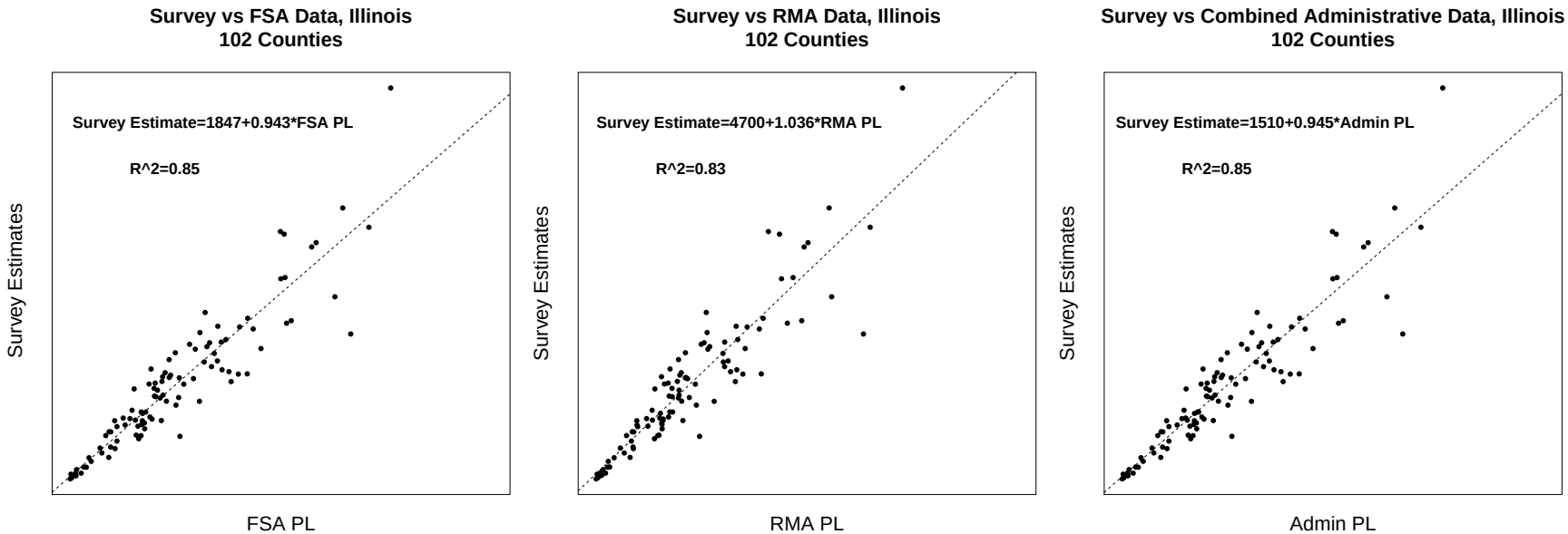
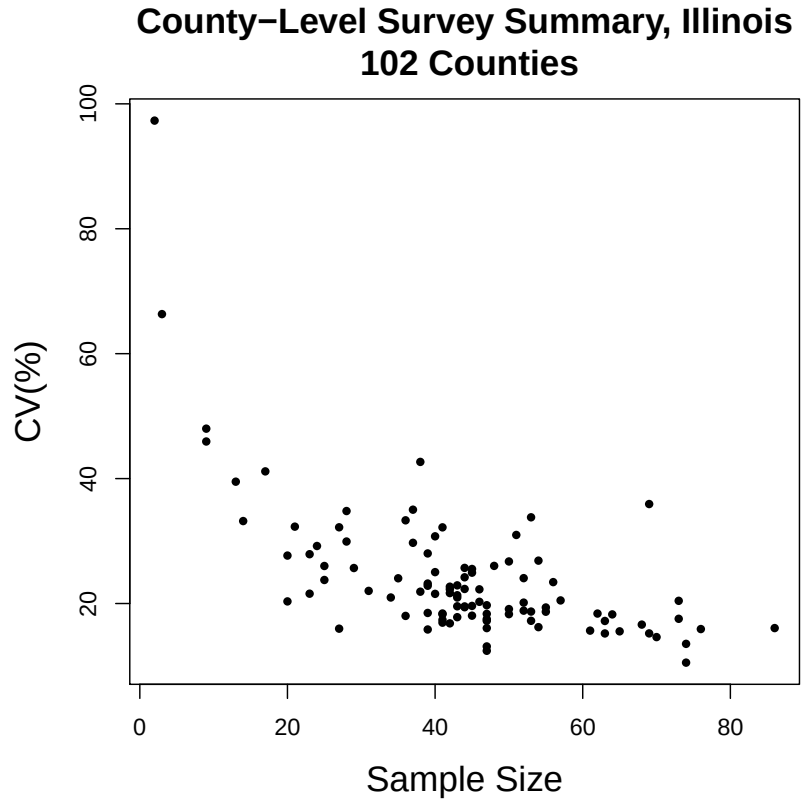


Table 2: Nationwide Summaries

	FSA PL			RMA PL			Admin 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	0.83	0.89	0.92

Admin PL: combine FSA and RMA, with preference for maximum planted acreage, available for 2401 counties

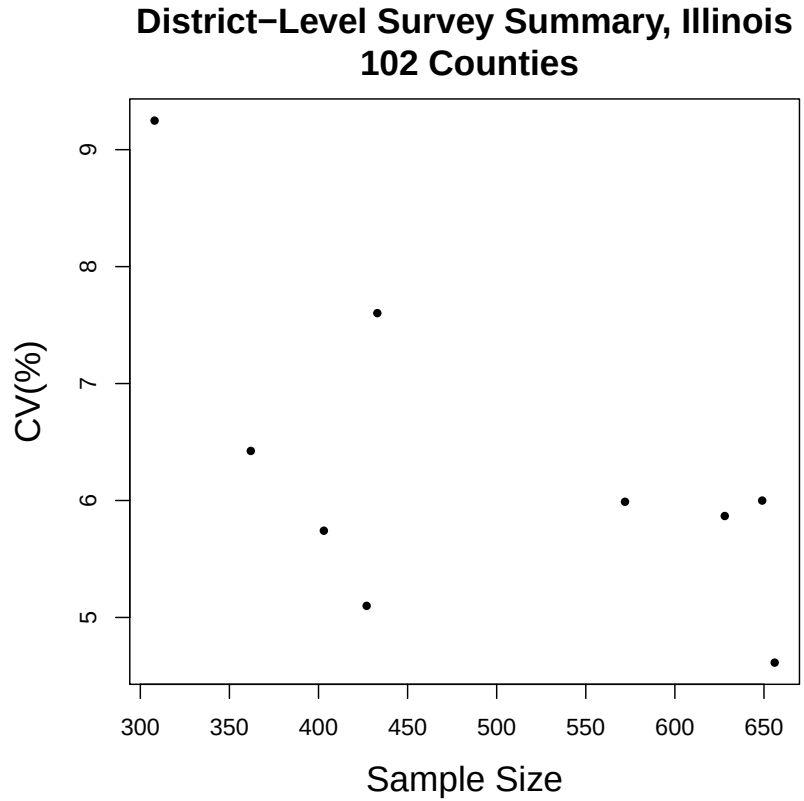
County-Level Relative Variability of Survey Estimates
2015 Corn Planted Acreage



Nationwide summaries

- sample size within a county: $[1, 191]$; median 18
- county-level CV(%): $[0.07, 107.66]$; median 31.94

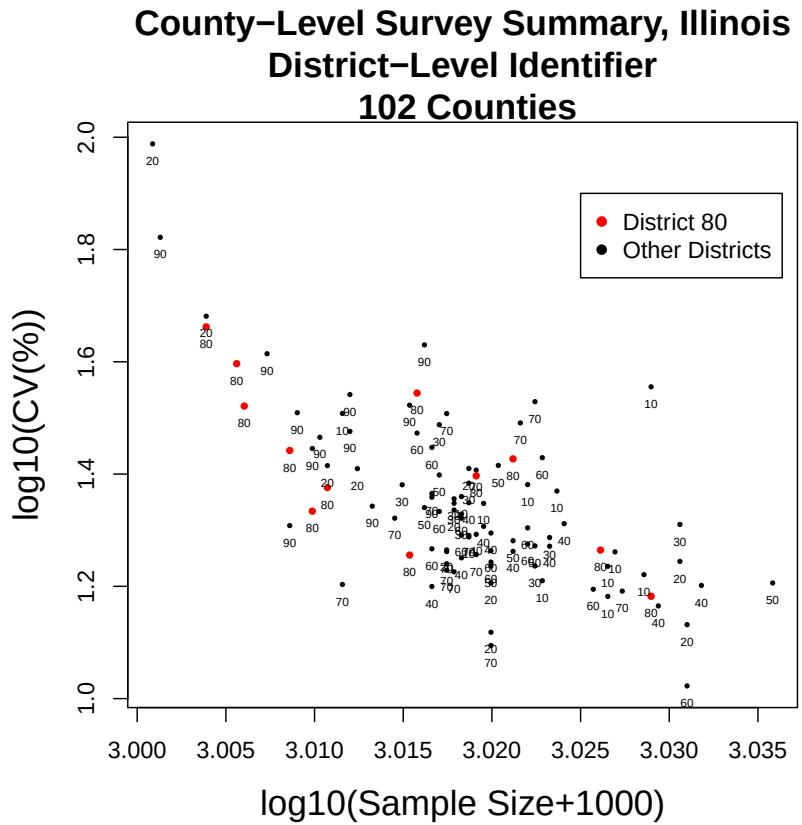
District-Level Relative Variability of Survey Estimates
2015 Corn Planted Acreage



Nationwide summaries

- sample size within a district: $[1, 993]$; median 145
- district-level CV(%): $[3.27, 100.70]$; median 11.84

Borrowing Information Across Counties and Districts
2015 Corn Planted Acreage



Nationwide summaries

- number of districts within a state: $[3, 15]$; median 9
- number of counties within a district: $[1, 32]$; median 8

Our Approach: Subarea-Level Model

Linkage model

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

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) in a given state
- $j = 1, ..., n_i^c$, subareas (counties) in area (district) i
- $\sum_{i=1}^{m} n_i^c = n^c$, number of counties in a given state
- 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

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 R 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: $\theta_{ij,r} \sim N(\mathbf{x}'_{ij}\boldsymbol{\beta}_r + v_{i,r}, \sigma^2_{u,r})$

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

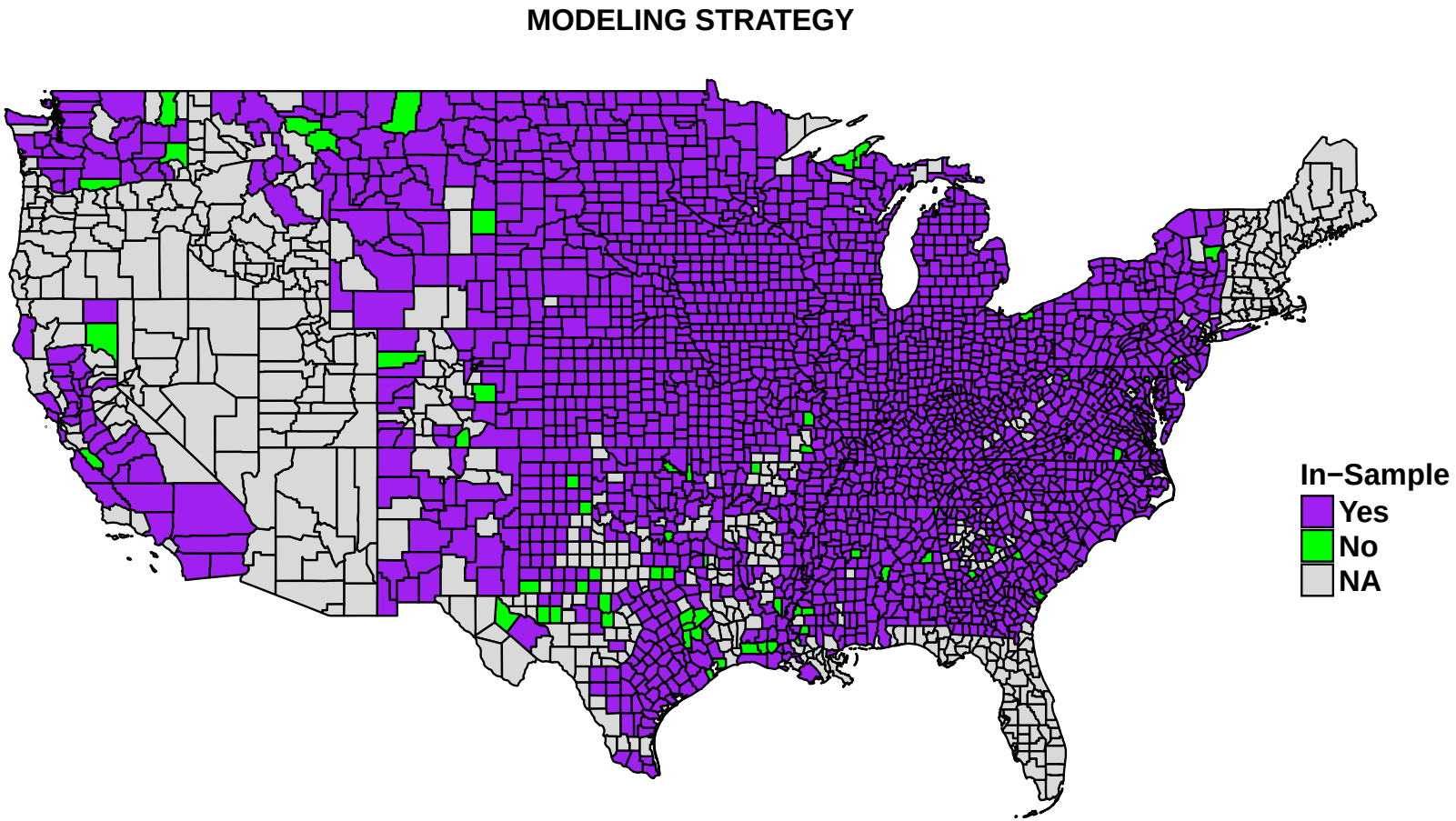
$$\theta_{ij,r}^{B} := \theta_{ij,r} \times a \times \left(\sum_{k=1}^{m} \sum_{l=1}^{n_k^{c*}} \theta_{kl,r} \right)^{-1},$$

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

Discussion:

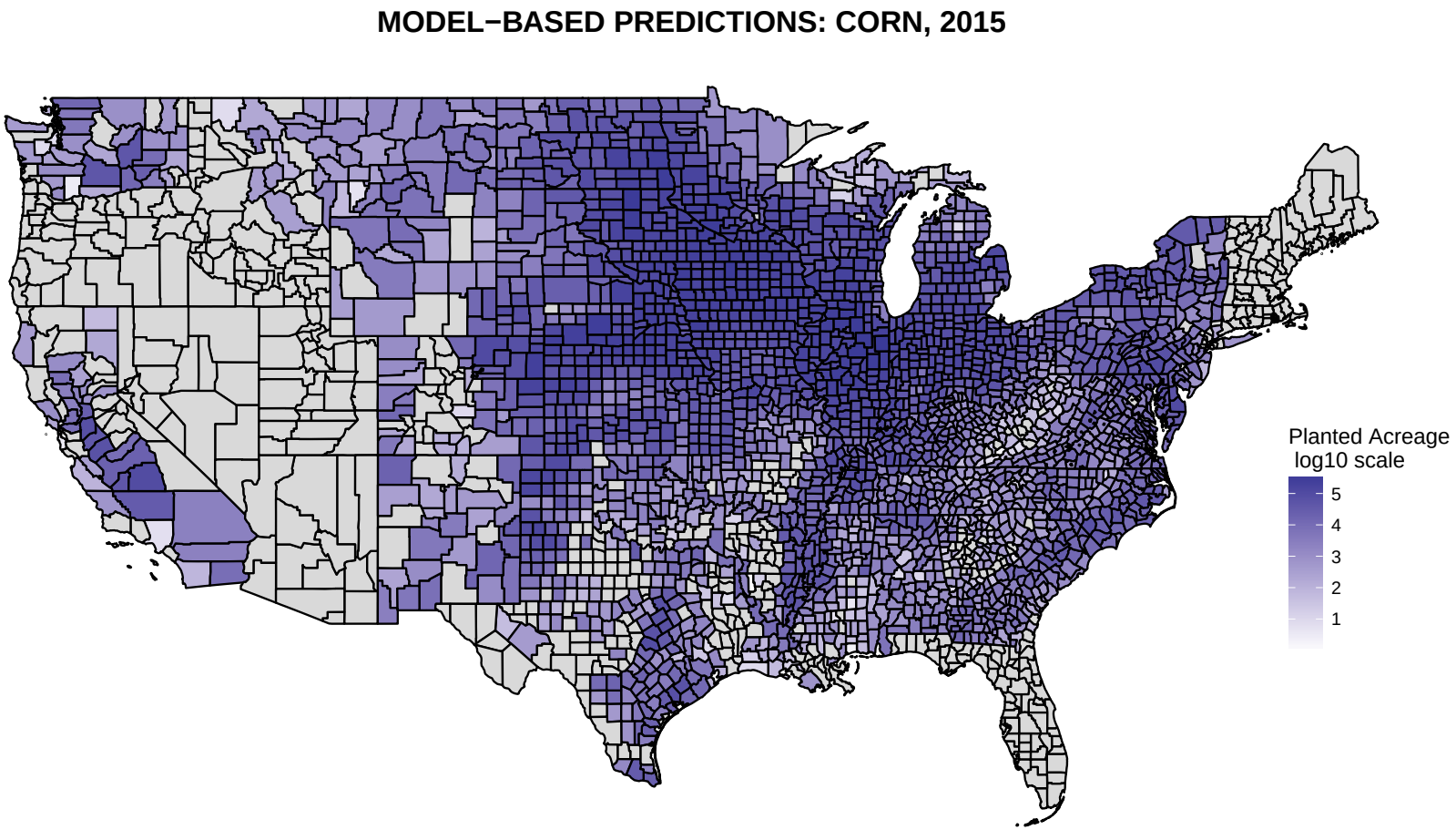
- defining the set of counties n^{c*}

Results



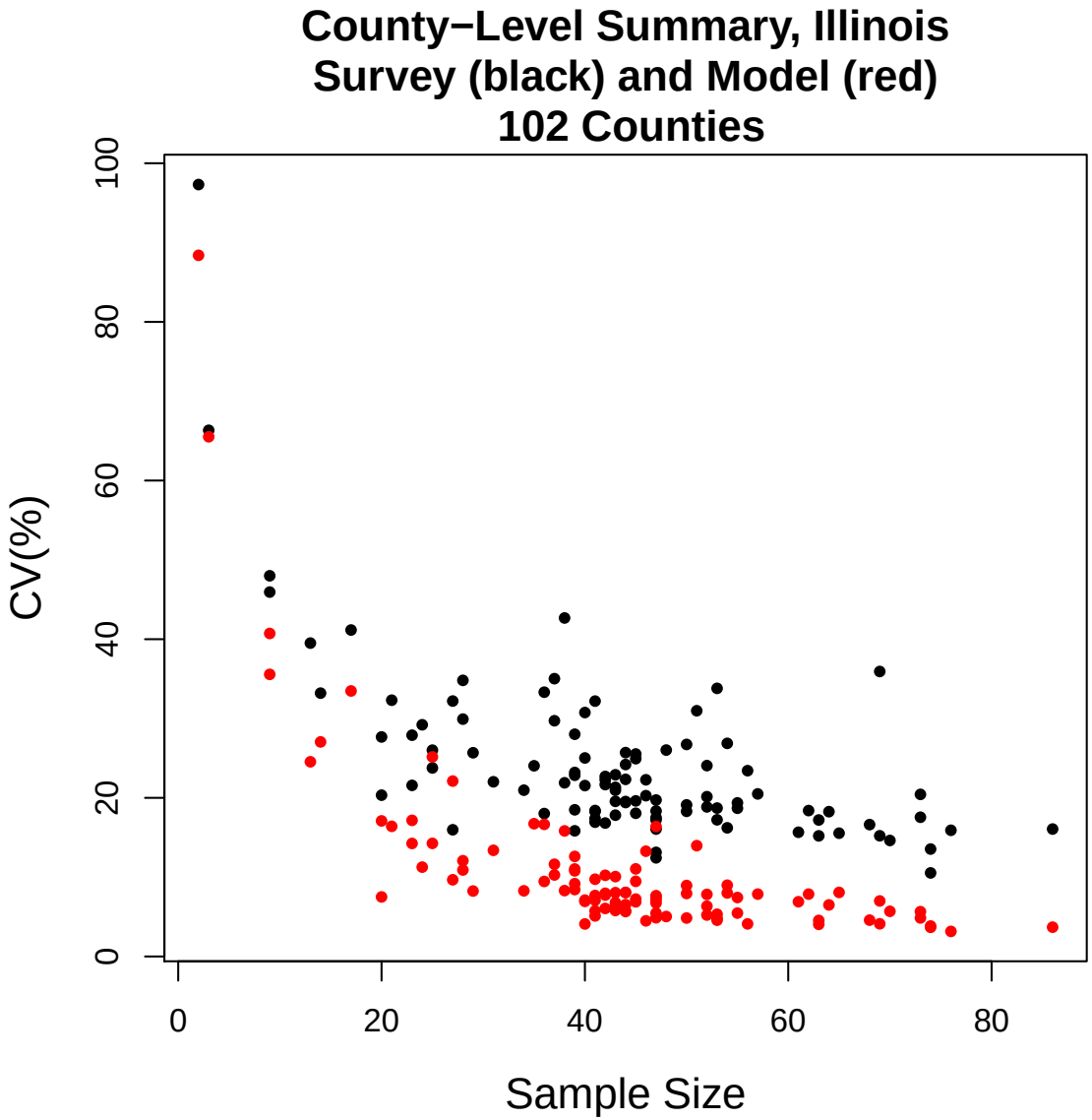
- 2423 in-sample counties and 70 not-in-sample counties

Results: Increased Number of County-Level Estimates

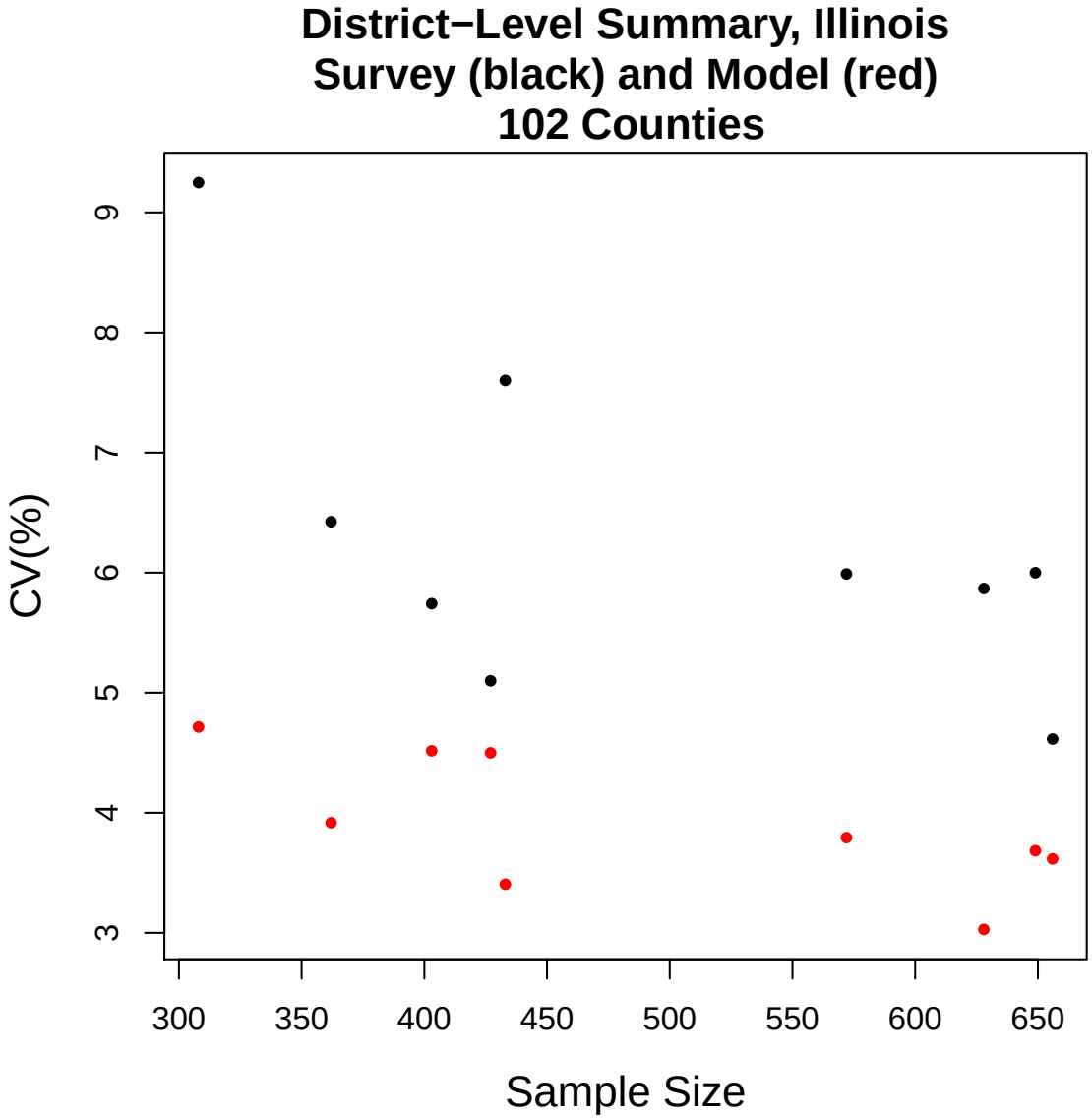


- model-based predictions available for 2493 counties
- RECALL: survey estimates available for 2426 counties

Results: Decreased Relative Variability



Results: Decreased Relative Variability



Results

Table 3: CV(%) summaries for counties/districts with available survey estimates

Level	Source	1st Qu.	Median	3rd Qu.
County	Survey	21.12	31.93	55.52
	Model	5.90	12.31	37.88
District	Survey	7.41	11.84	21.06
	Model	3.43	5.15	11.69

Discussion

- publication standard for official statistics
  - 2423 counties with available survey estimates:
    - 1125 survey CVs $\leq 30\%$ versus 1700 model CVs $\leq 30\%$
  - 2493 counties with available model-based predictions:
    - 1703 model CVs $\leq 30\%$
- 1622 counties published, under the current NASS publication standard; NASS QuickStats

Summary and Future Work

Summary

- model-based county-level and district-level predictions are produced, incorporating survey and administrative data $\Rightarrow$ increased number of county-level estimates
  - Texas: largest number of not-in-sample predictions, 20 out of 163 counties, accounting for $\sim$0.63% of planted acreage in the state
- reduction in precision and relative precision; model versus survey
  - $2 - 72\%$ / $19 - 74\%$ in most of the county-level SE / CV
  - $18 - 61\%$ / $28 - 66\%$ in most of the district-level SE / CV

Future work

- additional data sources $\Rightarrow$ revised set of counties to be estimated
- model specification; normality assumption
- quality of different data sources; imputation strategies
- publication standard

Thank you!

aerciulescu@niss.org

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., 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/.*

Results

Table 4: SE summaries for counties/districts with available survey estimates

Level	Source	1st Qu.	Median	3rd Qu.
County	Survey	640.50	2723.00	9464.00
	Model	428.70	1157.00	2848.00
District	Survey	4238.00	9242.00	31010.00
	Model	2106.00	5052.00	12360.00

Internal Model Validation
Posterior Predictive Checks

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

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

- 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