Model-based Estimation of the Number of U.S. Farms and Land in Farms from Survey and Administrative Data

Habtamu Benecha[1], Luca Sartore[2,1], Arthur Rosales[1], Grace Yoon[1], Tara Murphy[1], Denise Abreu[1]

[1]United States Department of Agriculture
National Agricultural Statistics Service (USDA NASS)

[2]National Institute of Statistical Science (NISS)

FCSM 2022 – Survey Research Methods Section

October 27, 2022

Disclaimer

The findings and conclusions in this presentation are those of the authors and should not be construed to represent any official USDA, or US Government determination or policy

- This research was supported in part by the intramural research program of the U.S. Department of Agriculture, National Agricultural Statistics Service (NASS)

PART I

MOTIVATION

The June Area Survey (JAS)

- The June Area Survey is one of the largest surveys conducted by NASS
- Based on an area frame, conducted every year during the first two weeks of June
  - All land in all states except Alaska stratified by land use
  - Land divided into homogeneous strata (e.g., cultivated land, urban area, etc.)
  - Each of the strata is divided into several substratum containing agriculturally similar areas
  - Each substratum is then divided into primary sampling units (PSU)
  - PSUs are sampled, divided into smaller areas called segments

The June Area Survey (JAS)

- One segment is randomly selected from each PSU

- Each segment is divided into several smaller areas called tracts
- Each tract is screened: agricultural or non-agricultural
- Agricultural tracts within selected segments are enumerated

Estimation from the JAS

- Annual estimates of the number of U.S. farms and land in farms are produced by using JAS data
- A farm is any place from which $1,000 or more of agricultural products were produced and sold or normally would have been sold during the year.
- Studies have shown that the number of U.S. farms derived from the JAS are underestimated, mainly due to misclassification during screening
  - A capture-recapture approach is employed to make estimation
  - NASS's list frame used as the second sample
  - JAS and list frame records are linked by using probabilistic record linkage
  - Different adjustment weights are estimated by using the matched dataset

Estimation of the number of farms and land in farms

For a farm tract i, the capture-recapture weight is estimated by

$$w_i = \frac{t_i}{\pi_i} \frac{Pr(F_i|S_iA_iR_iJ_i)}{Pr(J_i|S_iA_iR_iF_i)Pr(R_i|S_iA_iF_i)Pr(A_i|S_iF_i)},$$

where,

i = Tract on the JAS,
t_i = Proportion of a farm represented by tract
π_i = Sample inclusion probability
S_i = Tract is in the sample
A_i = Tract passed Agricultural screening process
R_i = Tract responds
J_i = Tract is recorded as a farm on the JAS
F_i = Tract is truly a farm

The Number of farms estimates are obtained by summing the capture-recapture weights over farm tracts

Estimation of the number of farms and land in farms

- A total of five logistic regression models
- Matched JAS & list-frame records are used
- Both agricultural & non-agricultural tracts used in the analysis
- Model covariates
  - JAS data available for agricultural tracts
  - JAS data not collected for non-agricultural tracts, challenging to get model covariates

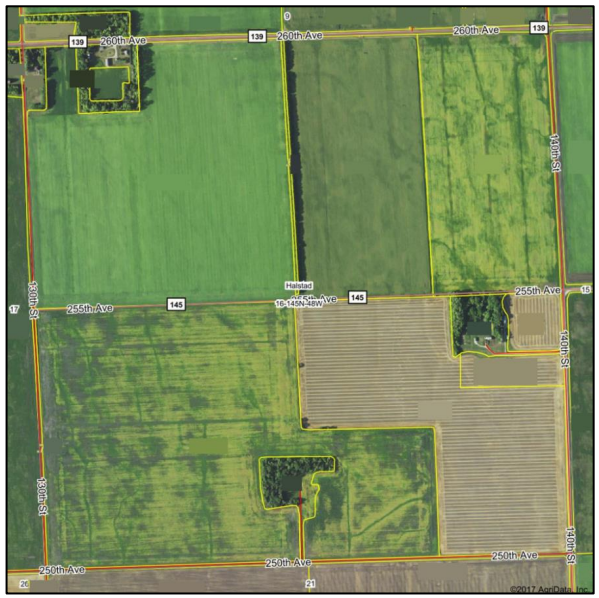
Goal: Using adminstrative & remote sensing data to create model covariates

PART II

PROPOSED METHODOLOGY

Farm Service Agency (FSA) data

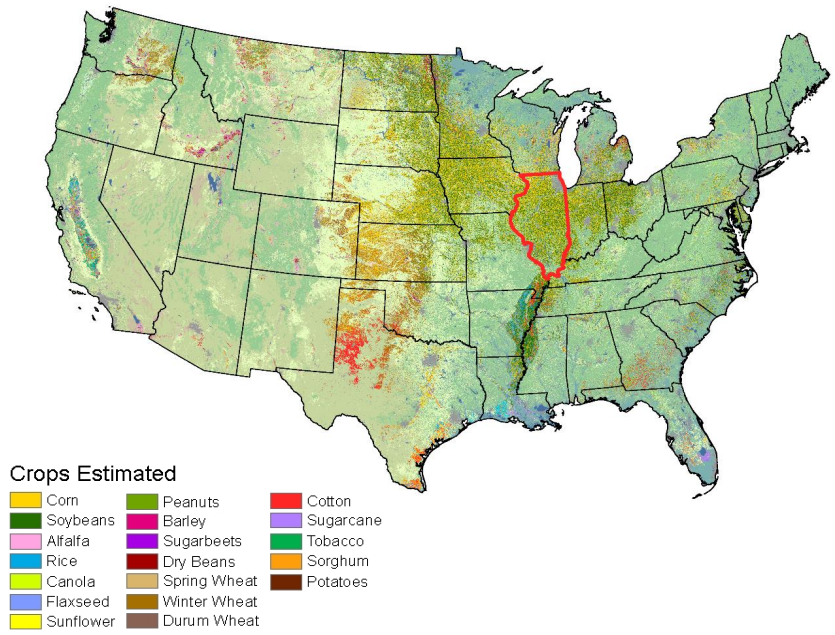
- FSA Common Land Units (CLUs)
  - The smallest unit of land that has a permanent contiguous boundary, a common land cover and land management, a common owner, and a common producer
- FSA-578 Form
  - Available for all land associated with a USDA program in a calendar year
  - Provides crop information



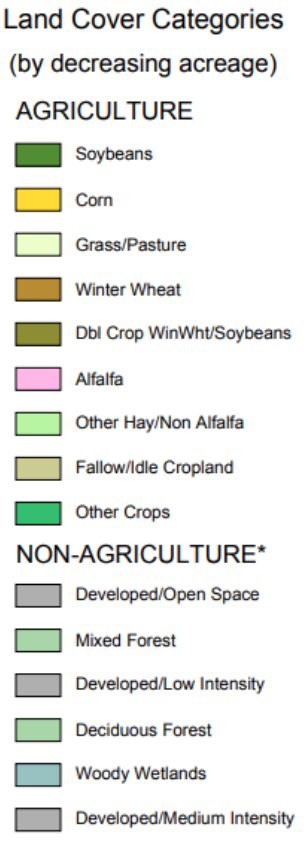
https://www.agridatainc.com/Home/Products/Mapping%20Features/Land%20Resource%20Intelligence/FSA%20Field%20Boundaries%20(CLU)

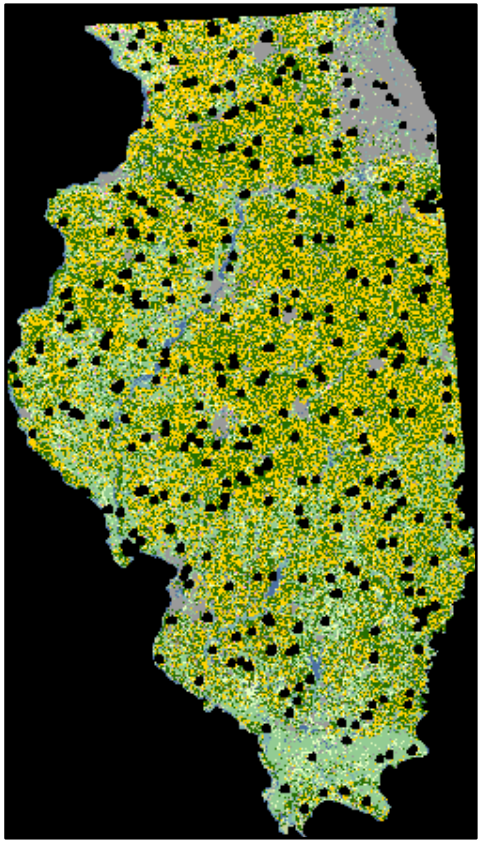
Cropland Data Layer (CDL)

- Georeferenced, crop-specific land cover raster dataset
- Captures planted acres
- National scale, 30m, 9 billion+ pixels
- Annually produced since 2008
- 85%-95% accuracy for major crops
- Freely available and open to the public

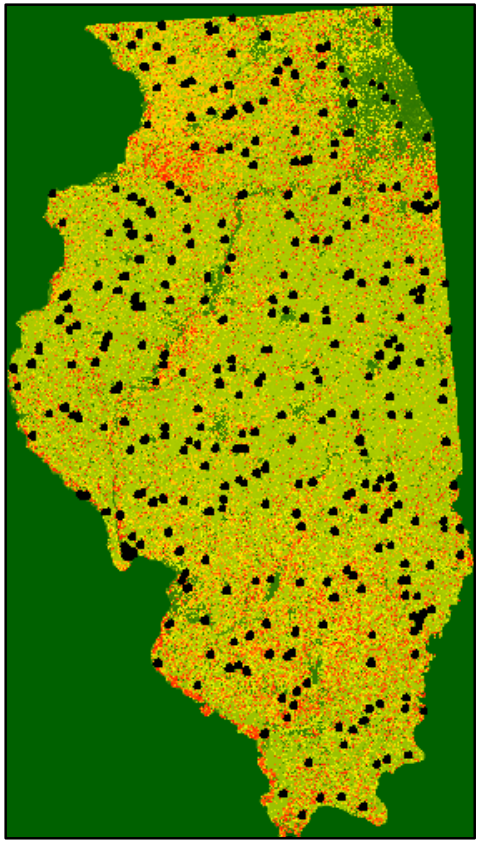


Predictive Cropland Data Layer (PCDL)

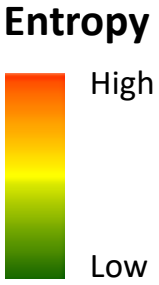




Illinois (2021)
PCDL and Segments



Illinois (2021)
Entropy Layer



PART III

APPLICATION

Case study

- Data available for 2021 have been considered in this study
- The JAS & digitized tracts data merged by using state, segment and tract identification numbers
- Tract-level PCDL data are used to create covariates for four models ($Pr(J_i|S_iA_iR_iF_i)$, $Pr(R_i|S_iA_iF_i)$, $Pr(A_i|S_iF_i)$, and farm status)
  - Predicted CDL for corn, cotton and rice
- Estimates of the number of farms and land in farms produced based on the capture-recapture model
  - State & U.S. level estimates produced without accounting for Alaska & Hawaii

Results

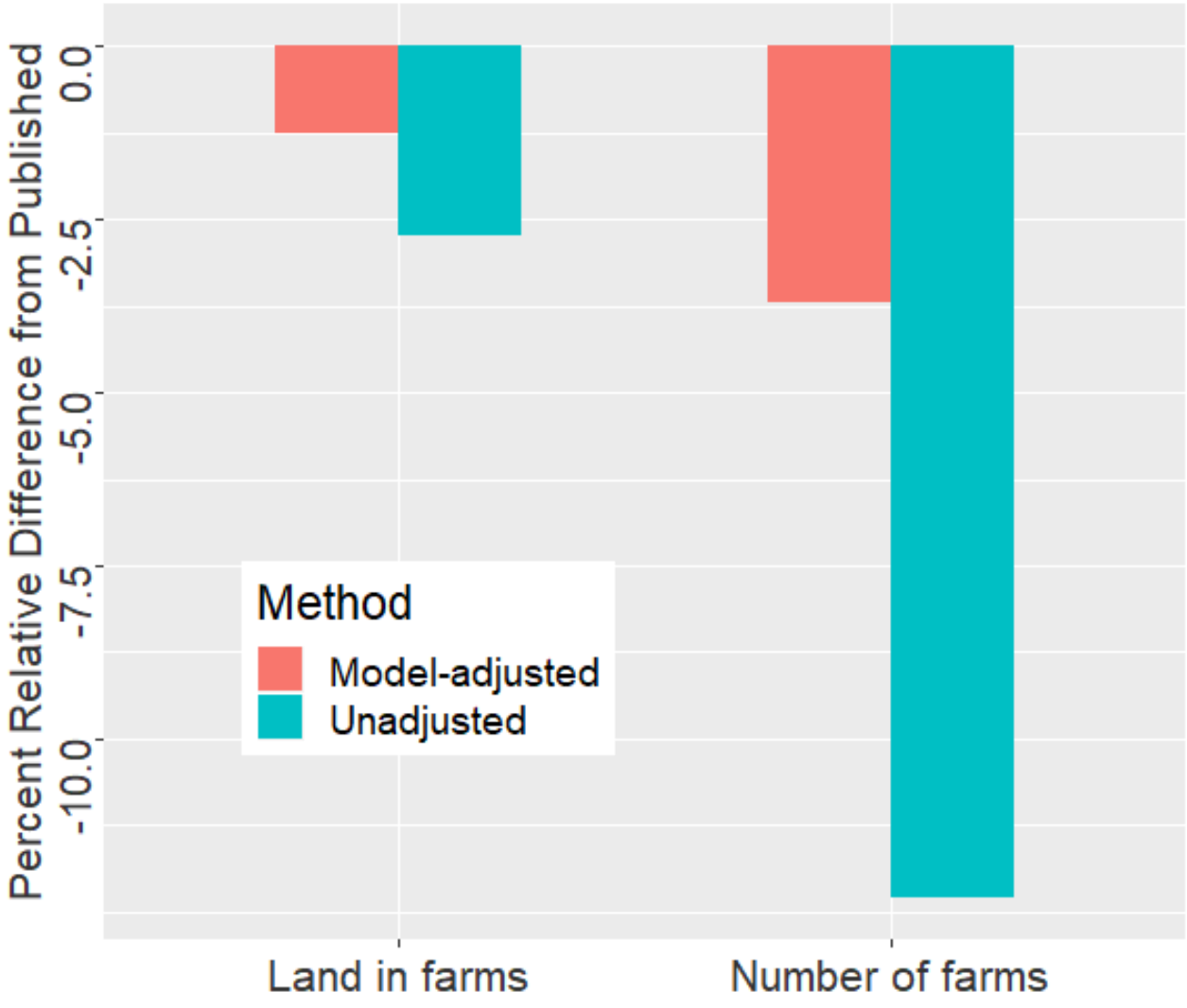


Figure: Percent relative differences of unadjusted and adjusted number of farms and land in farms against published (Year 2021)

PART IV

CONCLUSION

Final remarks

- The model-based adjustment improved state and national estimates
- Tract-level administrative and remote sensing data enabled to obtain more accurate adjustment weights
- Digitized tracts data can be used to obtain information on non-agricultural JAS tracts
- Challenges:
  - Complete digitized tracts data available only for 2021
  - Digitized tracts data not available for some of the JAS tracts
- Future research opportunities:
  - Summarizing the entropy layer at the tract level to be used as a covariate
  - Testing generalized additive models or other non-parametric models to better account for nonlinear effects

Select References

Abreu, D. A., McCarthy, J. S., Colburn, L. A., et al. (2010). Impact of the screening procedures of the June Area Survey on the number of farms estimates. *Research and Development Division. RDD Research Report Number RDD-10-03. Washington, DC: USDA, National Agricultural Statistics Service*.

Alho, J. M. (1990). Logistic regression in capture-recapture models. *Biometrics*, pages 623–635.

Boryan, C., Yang, Z., Mueller, R., and Craig, M. (2011). Monitoring us agriculture: the us department of agriculture, national agricultural statistics service, cropland data layer program. *Geocarto International*, 26(5):341–358.

Congalton, R. G. and Green, K. (2019). *Assessing the accuracy of remotely sensed data: principles and practices*. CRC press.

Hyman, M., Sartore, L., and Young, L. J. (2021). Capture–recapture estimation of characteristics of US Local Food Farms using a web-scraped list frame. *Journal of Survey Statistics and Methodology*.

Johnson, D. M. and Mueller, R. (2021). Pre-and within-season crop type classification trained with archival land cover information. *Remote Sensing of Environment*, 264:112576.

Mahammad, S. S. and Ramakrishnan, R. (2003). Geotiff-a standard image file format for gis applications. *Map India*, pages 28–31.

Osman, J., Inglada, J., and Dejoux, J.-F. (2015). Assessment of a markov logic model of crop rotations for early crop mapping. *Computers and Electronics in Agriculture*, 113:234–243.

Särndal, C.-E., Swensson, B., and Wretman, J. (2003). *Model assisted survey sampling*. Springer Science & Business Media.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.

USDA, FSA (2017). Farm Service Agency (FSA) Common Land Unit (CLU) information worksheet. `https://www.fsa.usda.gov/Assets/USDA-FSA-Public/usdafiles/APFO/support-documents/pdfs/clu_infosheet_2017_Final.pdf`.

Young, L. J., Lamas, A. C., and Abreu, D. A. (2017). The 2012 Census of Agriculture: a capture–recapture analysis. *Journal of Agricultural, Biological and Environmental Statistics*, 22(4):523–539.

Thank you!

Questions?

Habtamu Benecha, PhD
habtamu.benecha@usda.gov

Luca Sartore, PhD
luca.sartore@usda.gov