Automatic Imputation for an Area Survey

Tara Murphy[1], Arthur Rosales[1], Luca Sartore[1,2], Denise Abreu[1]

[1]National Agricultural Statistics Service, [2]National Institute of Statistical Sciences

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

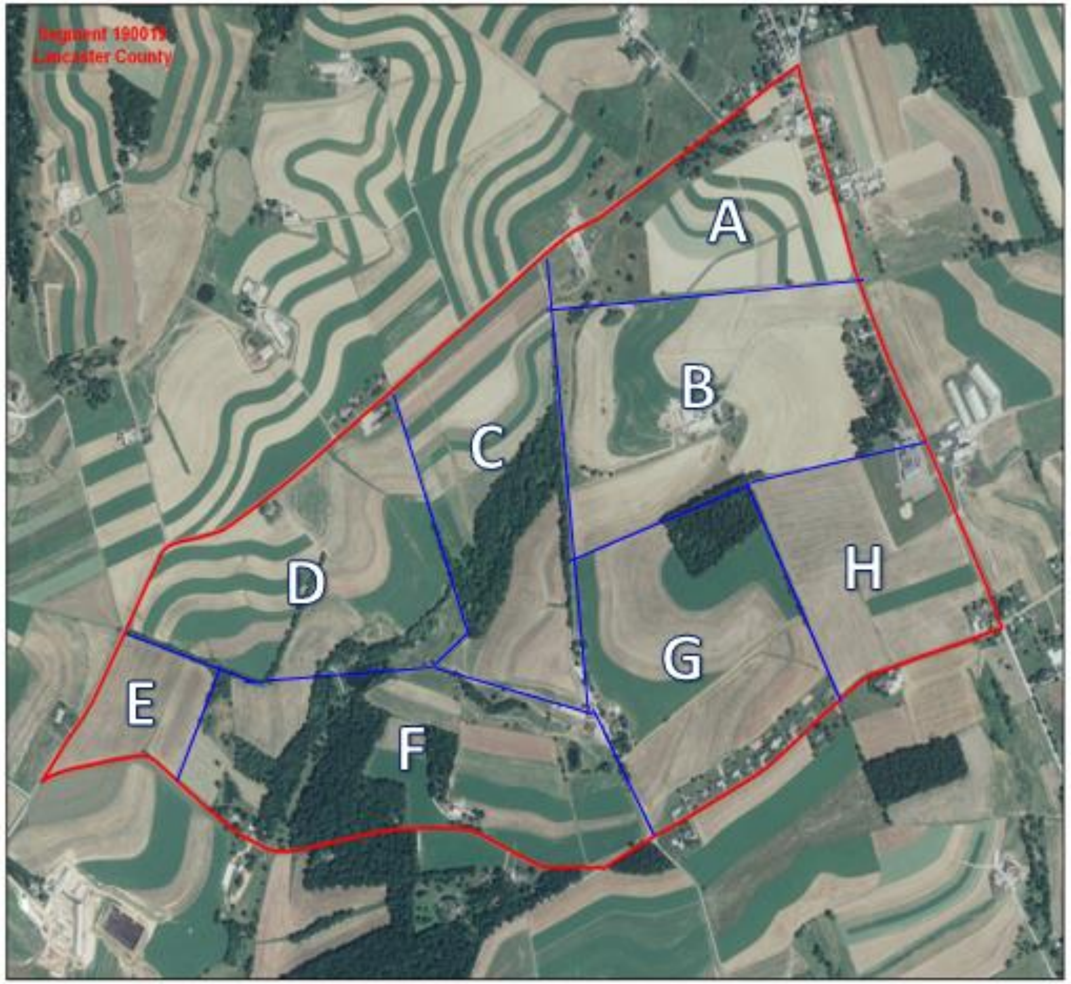
Background: June Area Survey (JAS)

- United States Department of Agriculture (USDA) National Agricultural Statistics Service's (NASS) largest annual survey

- Provides key indications for many agricultural aspects, including:
  - Planted acreage for most row crops and small grains

- Measures the incompleteness of the NASS List Frame

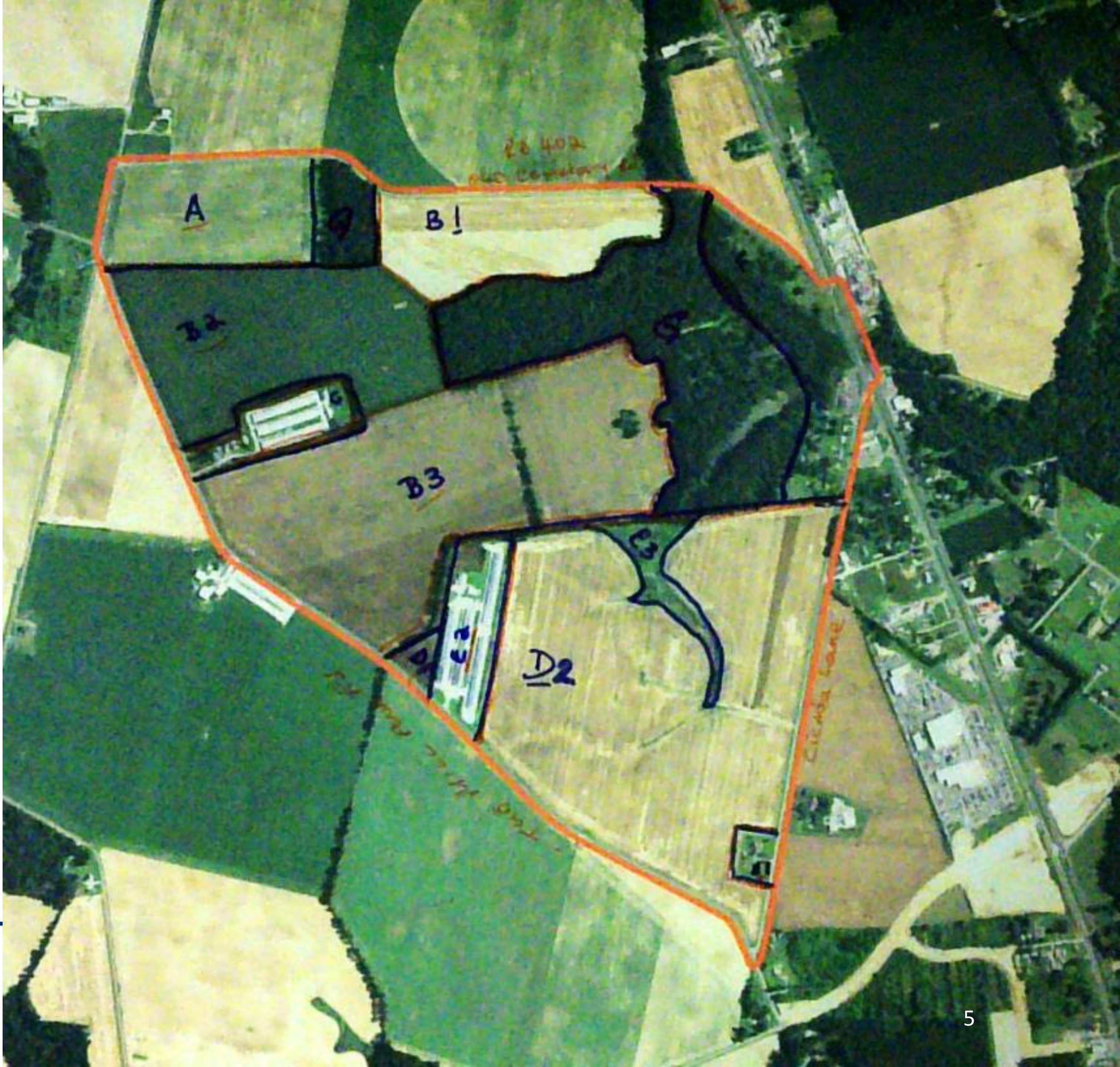
Background: June Area Survey (JAS)



- Area-frame based
- Segments of land sampled and remain in survey for five years
- Sampled segments divided into tracts representing unique land operating arrangements

Unique Nonresponse Challenges for JAS Tracts

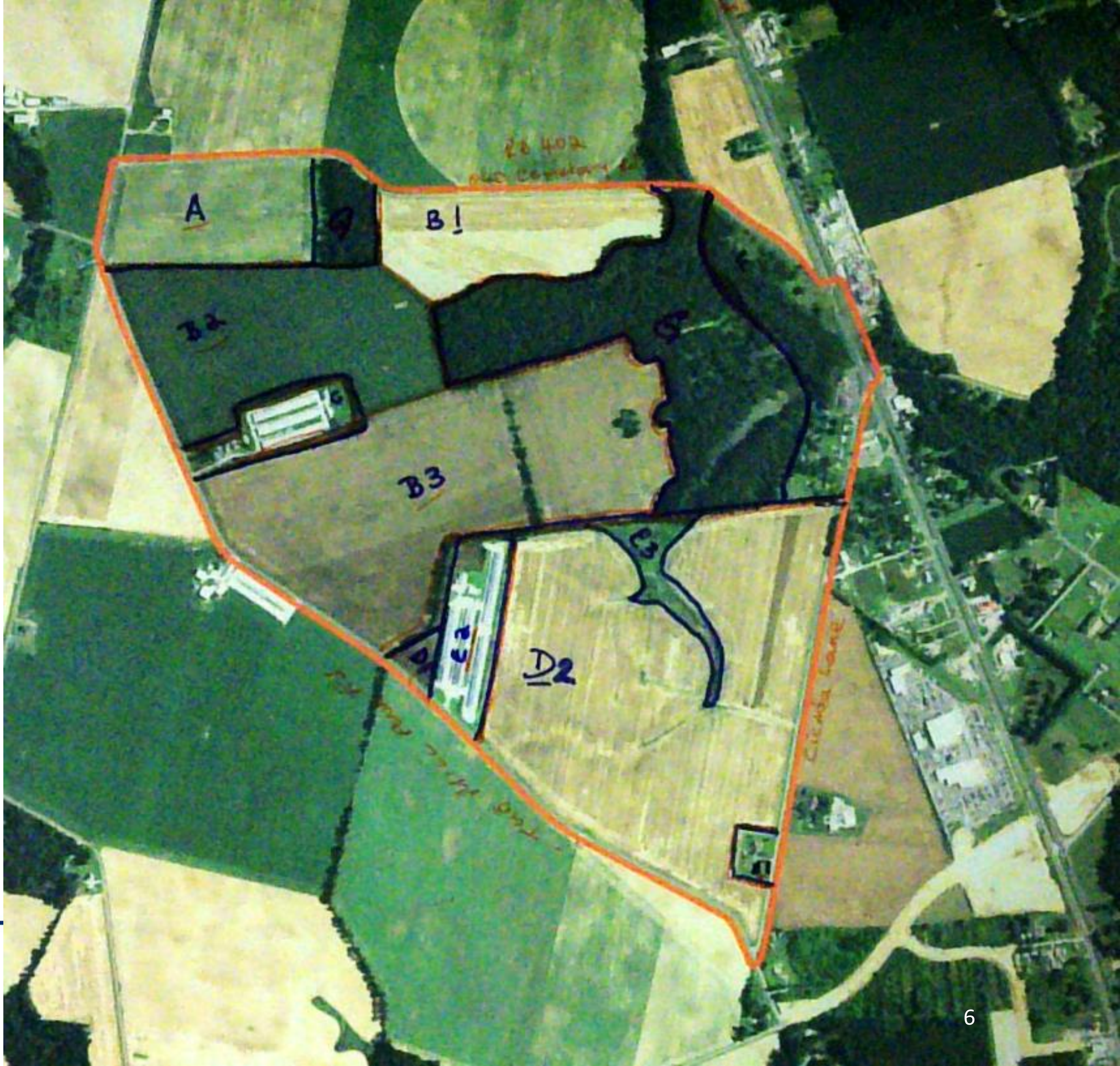
- Data collection based mostly on in-person and telephone interviews
- Extensive screening activities are needed to identify in-scope land tracts, especially for new segments



Unique Nonresponse Challenges for JAS Tracts

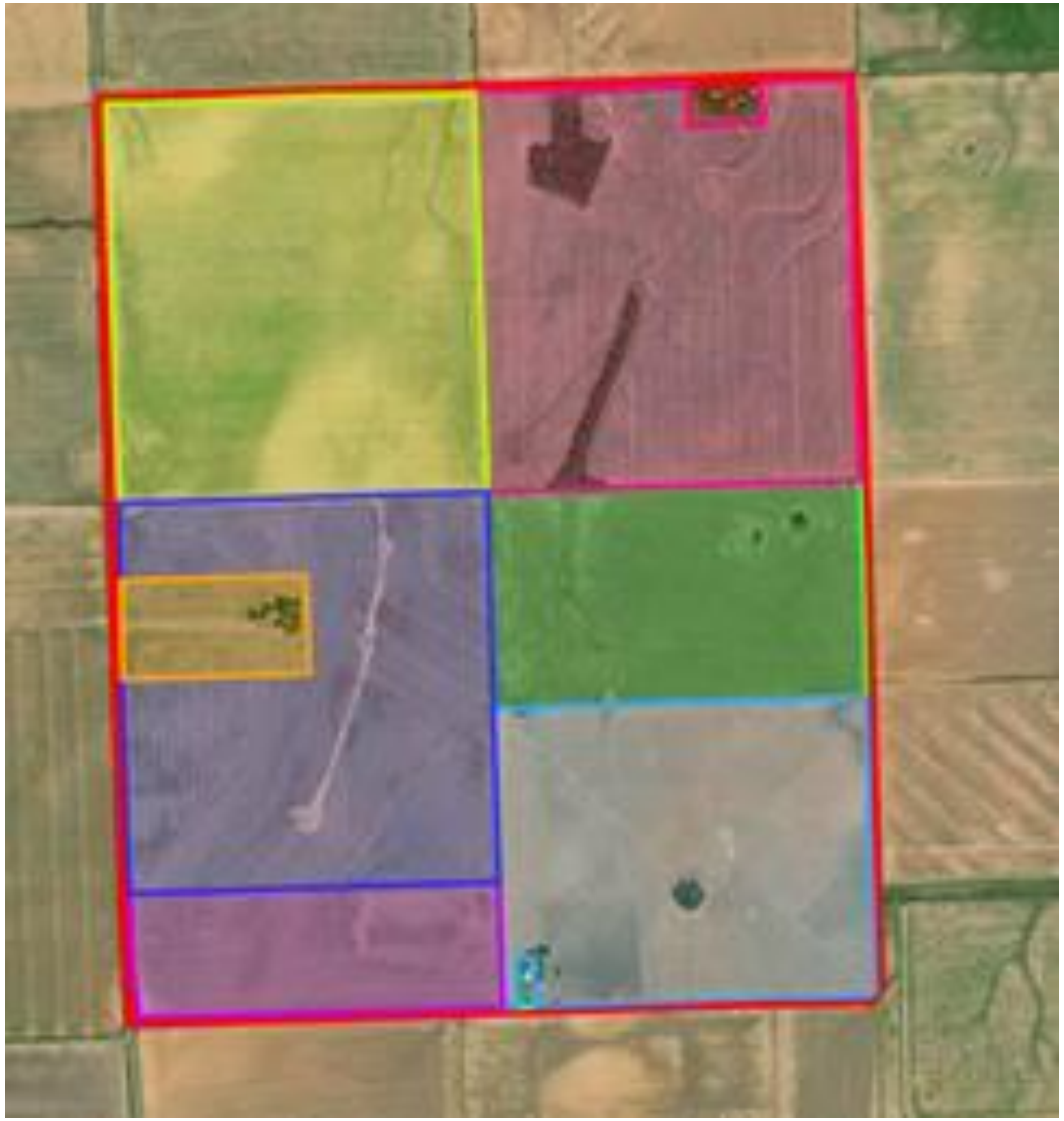
- Land-use arrangements may change during the five-year sample period

- **Digital records of tract boundaries** have historically never been created, making it difficult to link external, ancillary data to those tracts



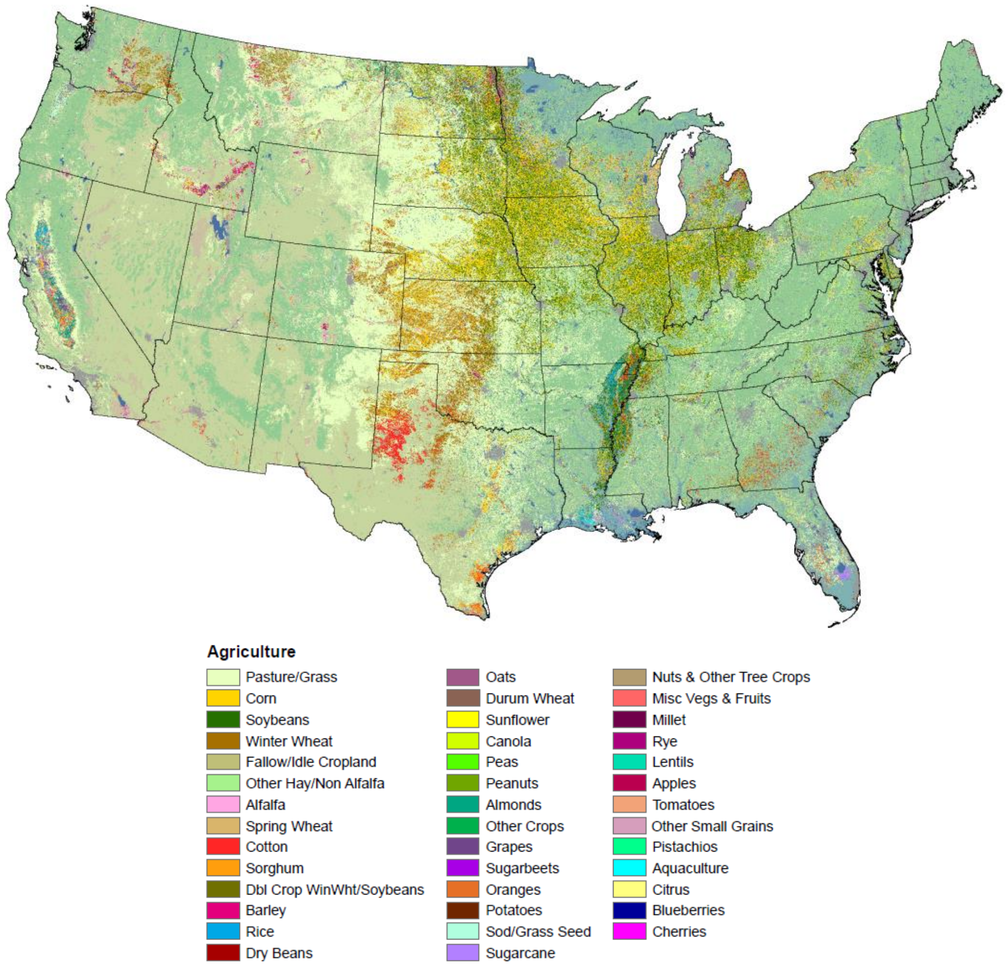
Availability of New Data

- Beginning 2021, digitization of all in-sample tract boundaries performed
- Resulting in geospatially-referenced record of tracts
- Allowing them to be linked to other data for estimation and imputation



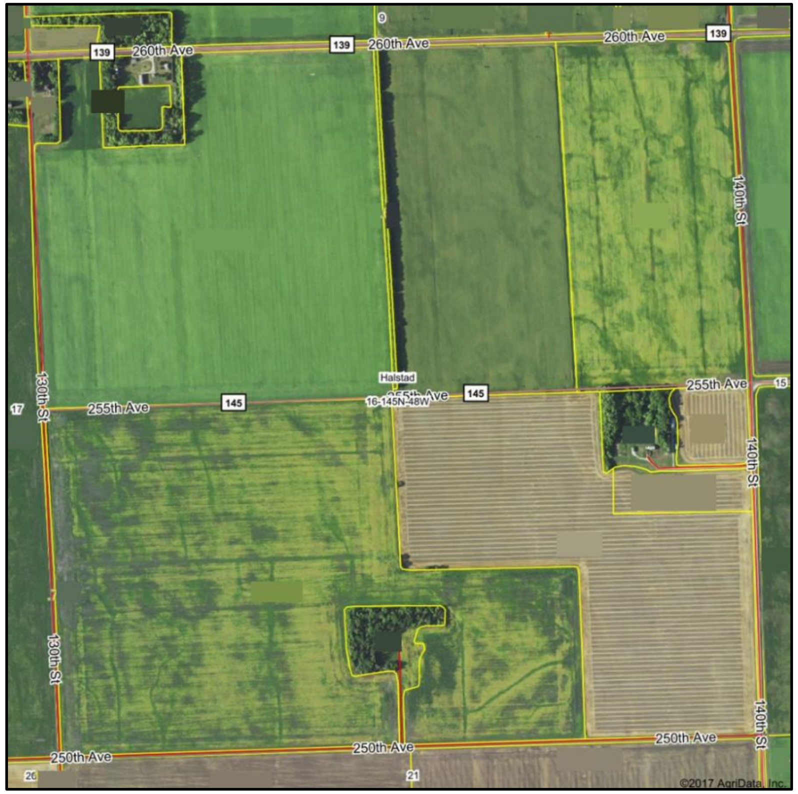
Auxiliary Data: Cropland Data Layer (CDL)

- Crop-specific land cover classification product created by USDA NASS
  - Raster product at 30-meter resolution
  - Available for the conterminous U.S. annually since 2008
  - Pixel-level crop data for over 100 crop categories
  - Only available at the end of the year



Auxiliary Data: Farm Service Agency (FSA)

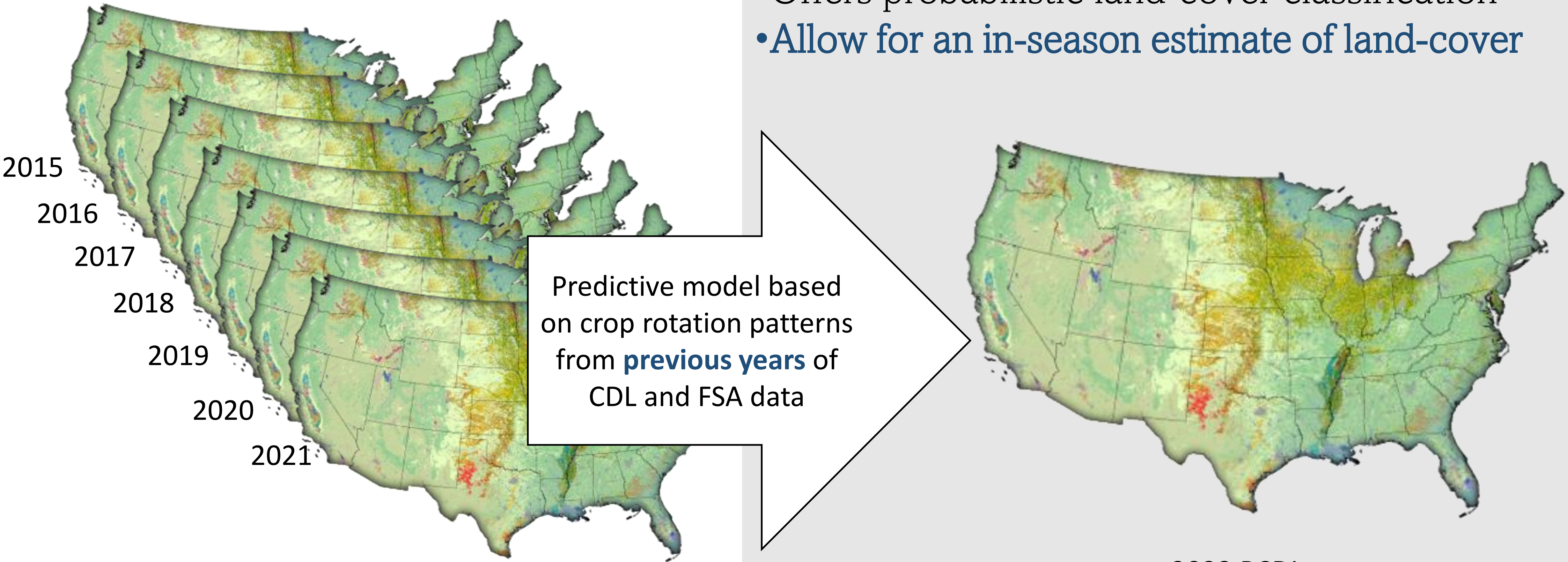
- FSA-578 Form
  - Available for all land associated with a USDA program in a calendar year
  - Provides crop information (what producers are growing and where)
- FSA Common Land Units (CLUs)
  - Smallest unit of land that has a permanent contiguous boundary, common land cover and land management, common owner, & common producer
  - CLUs linked to corresponding FSA-578 data



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

New Auxiliary Data: Predictive CDL (PCDL)

- Offers probabilistic land-cover classification
- Allow for an in-season estimate of land-cover



Predictive model based on crop rotation patterns from **previous years** of CDL and FSA data

2022 PCDL

Entropy Layer of PCDL

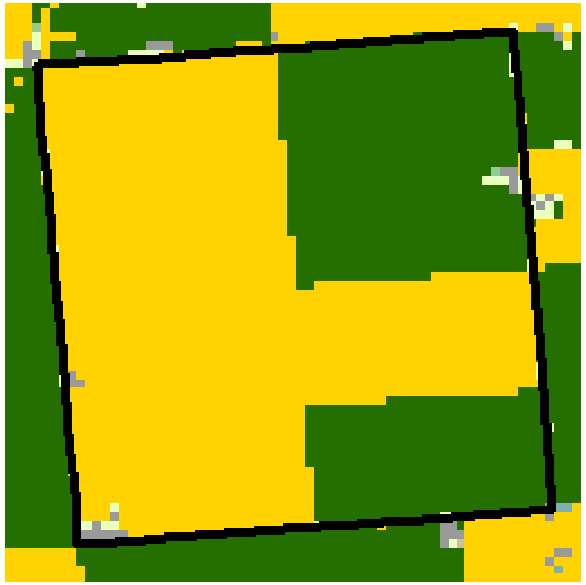
- Designed to provide a sense of confidence in the PCDL for the area of interest

- Low entropy = High predictability
- High entropy = Low predictability

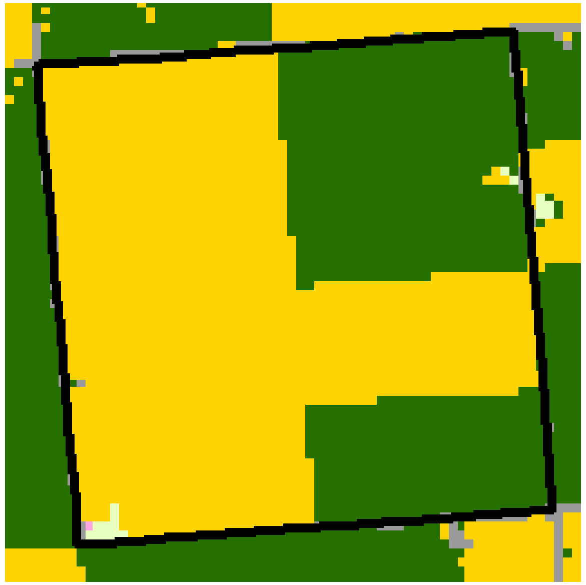
Entropy Layer of PCDL

🟨 Corn

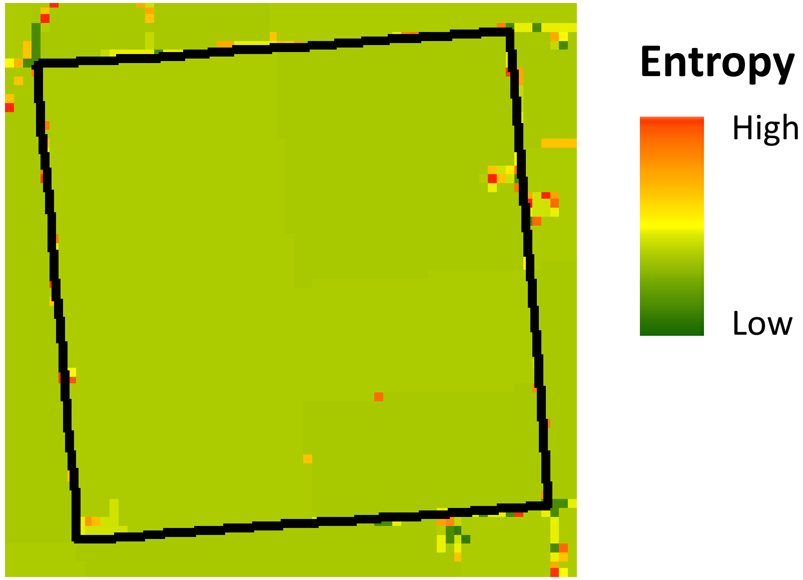
🟩 Soybeans



FSA CLUs linked to 578 data & Official CDL “Truth”

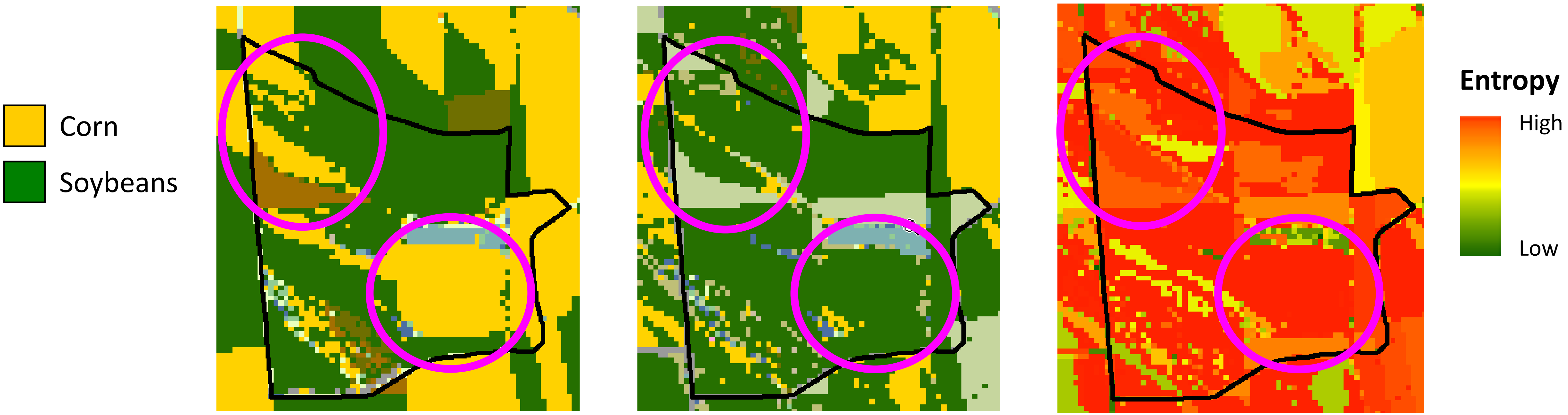


Predictive CDL



Entropy Layer Low Entropy

Entropy Layer of PCDL



FSA CLUs linked to 578 data & Official CDL "Truth"

Predictive CDL

Entropy Layer
High Entropy

Research Question

- Can automatic imputation be performed at the tract-level by incorporating digitized tracts, the PCDL and the Entropy Layer?

Data Preparation

- 2019 & 2021 JAS survey data utilized
- PCDL and FSA data summarized within digitized tract boundaries and linked to respective JAS survey data
- Data was subset to “low hanging fruit” records:
  - Number of PCDL crops < 2
  - Mean entropy of tract < 0.1 (Sartore, et al., 2022)
  - Digitized tract acres between 10 and 1000 acres

Imputation Model

- Cubist model
  - Used to predict FSA crop acreage at the crop level, based on PCDL, entropy, and other covariates (e.g., lat/long, state, sampling stratum)
  - Implemented using *caret cubist* packages in R software

- Model fit on 2019 data
- Predictions made on 2021 data

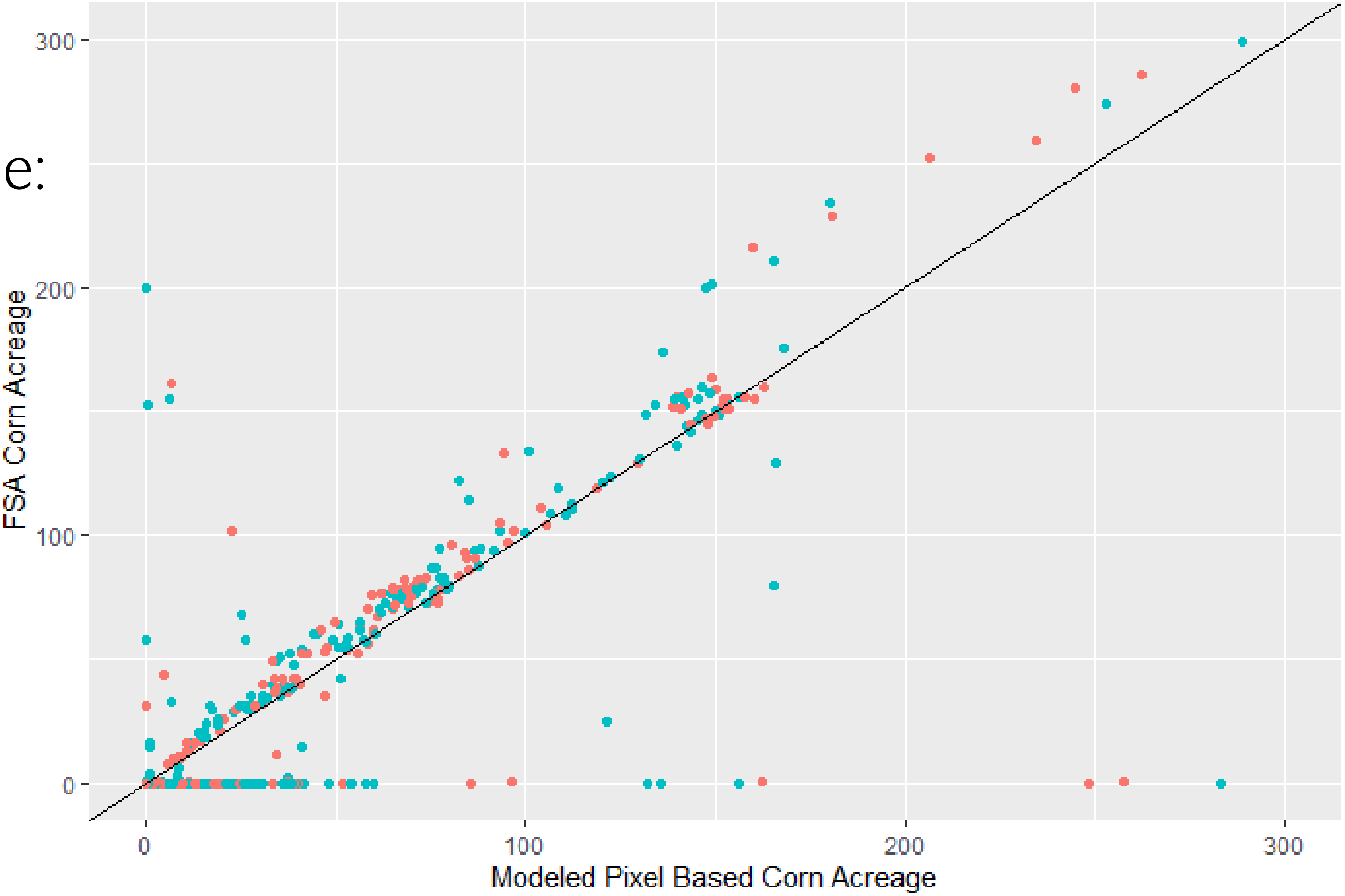
Results: Corn

- Color coded by JAS response:

●	0 = manually estimated
●	1 = reported

- $R^2 = 0.781$
- MAE = 4.783 acres
- Important model variable:
  - PCDL Corn

Model vs FSA Data

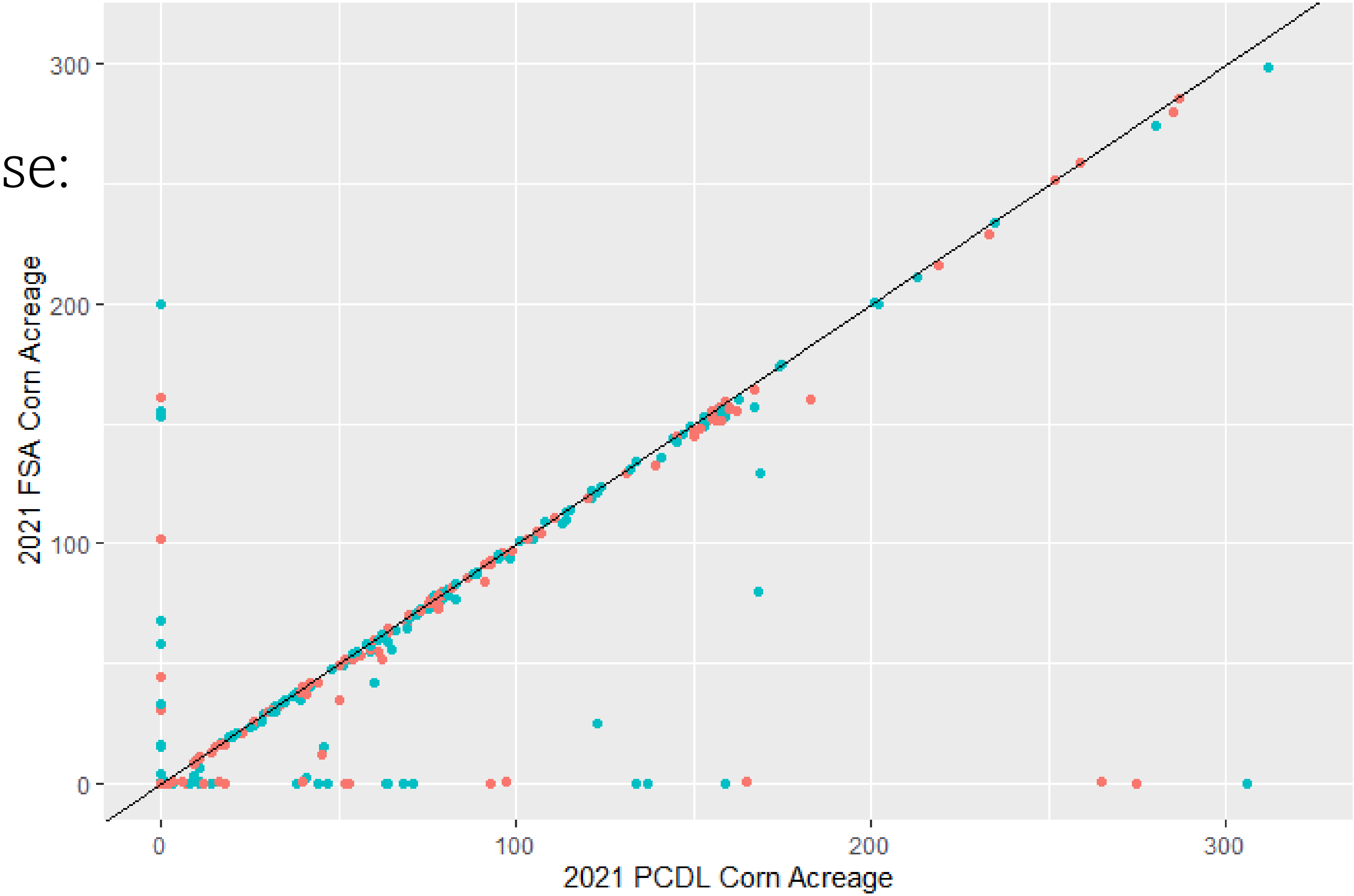


Results: Corn

- Color coded by JAS response:
  - ● 0 = manually estimated
  - ● 1 = reported

- $R^2 = 0.807$
- MAE = 2.797 acres

PCDL vs FSA Data



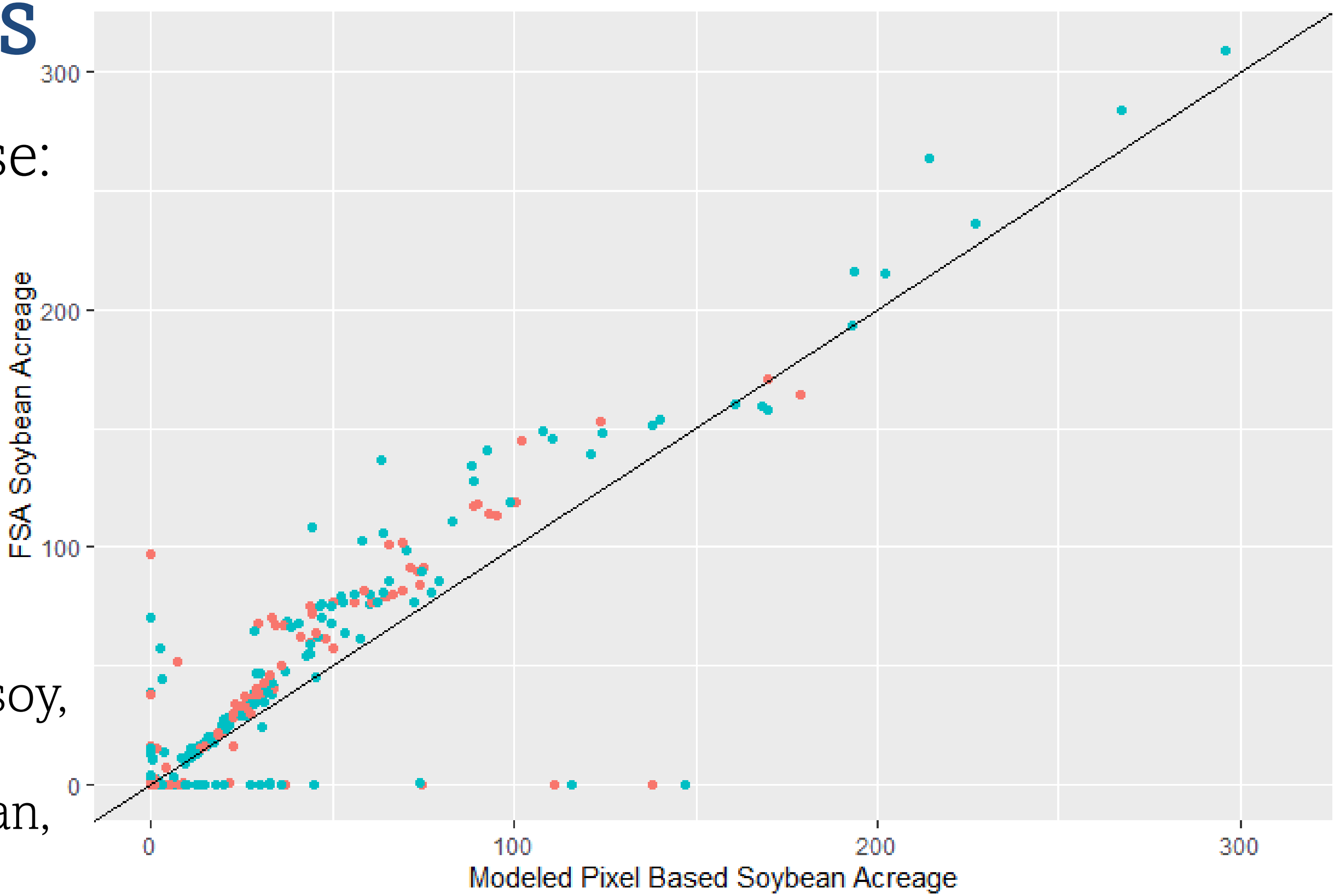
Results: Soybeans

- Color coded by JAS response:

  - ● 0 = manually estimated
  - ● 1 = reported

- $R^2 = 0.86$
- MAE = 2.82 acres
- Important model variables:
  - PCDL combined soy, PCDL soy, digitized tract acres, state, entropy mean, entropy median, latitude, longitude

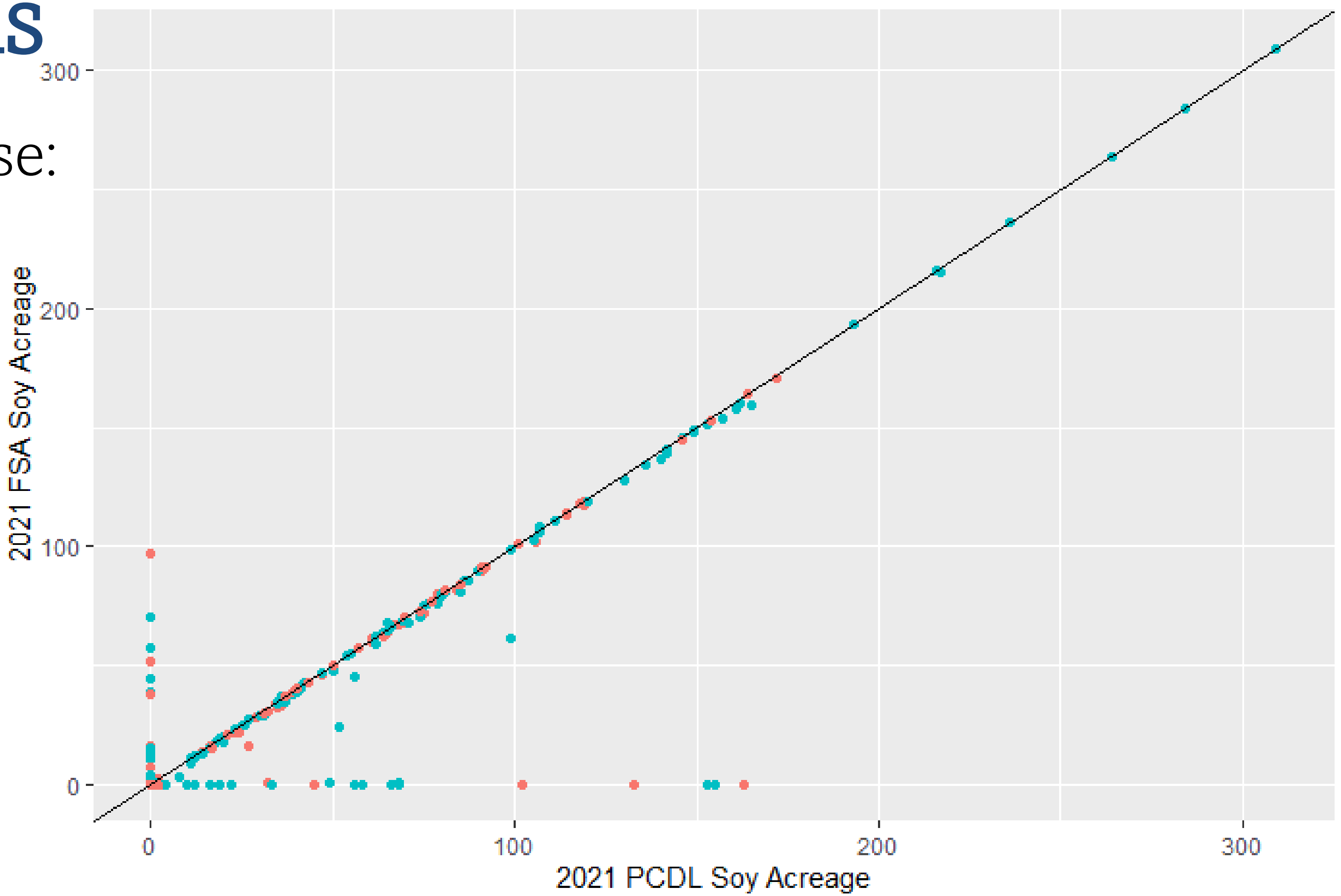
Model vs FSA Data



Results: Soybeans

- Color coded by JAS response:
  - ● 0 = manually estimated
  - ● 1 = reported
- $R^2 = 0.884$
- MAE = 1.336 acres

PCDL vs FSA Data



Discussion

- PCDL outperforms imputation model for “low hanging fruit”
  - Automatic imputation can easily be performed

- However, “low hanging fruit” represents a small portion of records

Future Research

- Expand beyond "low hanging fruit" records
- Find optimal level of entropy where PCDL is useful for the purposes of imputing JAS tract nonresponse
- Improve imputation model by incorporating additional auxiliary data
  - Economic data
  - Environmental data

Select References

Boryan, Claire, et al. "Monitoring US agriculture: the US department of agriculture, national agricultural statistics service, cropland data layer program." *Geocarto International* 26.5 (2011): 341-358.

FSA (2021) . Common Land Units (CLUs). < https://www.fsa.usda.gov/programs-and-services/aerial-photography/imagery-products/common-land-unit-clu/index>

Gerling, Michael W., HoaiNam N. Tran, and Terry P. O'Connor. *The Road to Understanding Nonresponse in the June Area Survey*. No. 1496-2016-130667. 2010.

Kolmogorov, A. (1956). On the Shannon theory of information transmission in the case of continuous signals. *IRE Transactions on Information Theory*, *2*(4), 102-108.

Lopiano, Kenneth K., et al. "Adjusting the June Area Survey for Non-response and Misclassification." *Proceedings of the Joint Statistical Meetings*. 2010.

Mueller, Rick, and Robert Seffrin. "New methods and satellites: a program update on the NASS cropland data layer acreage program." *Intl. Archives Photogrammetry, Remote Sensing, and Spatial Information Sci* 36 (2006): 8.

L. Sartore, C.G. Boryan, and P. Willis, “Developing Entropies of Predictive Cropland Data Layers for Crop Survey Imputation”. Proc. of IEEE International Geoscience and Remote Sensing Symposium (IGARSS), 2022 IEEE International, Kuala Lumpur, Malaysia, July 17 – 22, 2022

Shannon, C. E. (1951). Prediction and entropy of printed English. *Bell system technical journal*, *30*(1), 50-64.

Thank you!

Tara.Murphy@usda.gov