Data-fusion approaches to improve acreage estimates

Luca Sartore[1,2], Jake Abernethy[2]
Claire Boryan[2], Lu Chen[1,2], David M. Johnson[2]
Kevin A. Hunt[2], Clifford Spiegelman[3], Linda J. Young[2]

[1]National Institute of Statistical Science (NISS)
[2]United States Department of Agriculture
National Agricultural Statistics Service (USDA NASS)
[3]Texas A&M University, College Park (TAMU)

lsartore@niss.org

Extreme Machine Learning Methods and Applications
August 4, 2020

Disclaimer and acknowledgments

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).

A special thanks to Bob Seffrin, who produced acreage estimates for the case study

Presentation outline

1. Data fusion definition
2. NASS acreage report
3. Processing data sources
4. Case study
5. Concluding remarks

PART I

DATA FUSION DEFINITION

Definition

Many fusion methods exist in different scientific fields. In particular, data fusion blends information from several sources to provide more consistent and efficient results (Zhang, 2010)

This can be achieved through a data integration system that combines data from different sources (Lenzerini, 2002)

PART II

NASS ACREAGE REPORT

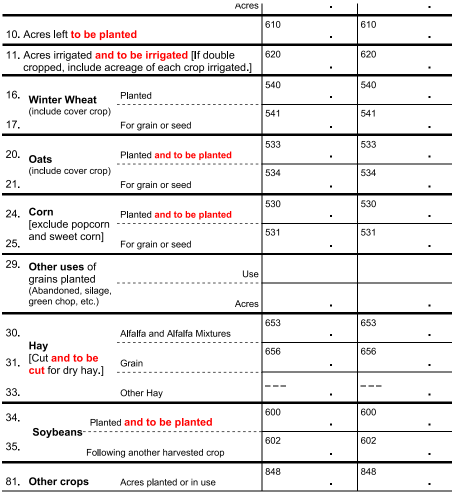
USDA NASS acreage report

USDA NASS publishes state and county level estimates of planted acreage by combining several sources of information

NASS publishes its acreage estimates at the state level several times during the year:

1. Prospective Planting (end of March)
2. Acreage (end of June)
3. Crop Production (monthly from August to December)

NASS questionnaire



		Acres	.	.
10.	Acres left **to be planted**		610 .	610 .
11.	Acres irrigated **and to be irrigated** [If double cropped, include acreage of each crop irrigated.]		620 .	620 .
16.	**Winter Wheat** (include cover crop)	Planted	540 .	540 .
17.		For grain or seed	541 .	541 .
20.	**Oats** (include cover crop)	Planted **and to be planted**	533 .	533 .
21.		For grain or seed	534 .	534 .
24.	**Corn** [exclude popcorn and sweet corn]	Planted **and to be planted**	530 .	530 .
25.		For grain or seed	531 .	531 .
29.	**Other uses** of grains planted (Abandoned, silage, green chop, etc.)	Use		
		Acres	.	.
30.	**Hay** [Cut **and to be cut** for dry hay.]	Alfalfa and Alfalfa Mixtures	653 .	653 .
31.		Grain	656 .	656 .
33.		Other Hay	--- .	--- .
34.	**Soybeans**	Planted **and to be planted**	600 .	600 .
35.		Following another harvested crop	602 .	602 .
81.	**Other crops**	Acres planted or in use	848 .	848 .

- June Area Survey
- June 1 reference date
- Two-week data collection
- Respondents also report intentions (‘to be planted’)

Intentions may change...

Heavy rains in 2019 affected planting activities

Effect on reported estimates

Heavy rains impacted subsequent planting activity

- User interest in planted area totals published June 28, 2019
- Announced re-contact efforts[1] with release of *Acreage* report

State	Corn *2018 Final* (1,000 Acres)	 *2019 June*[2] (% Change)	 *2019 August*[3] (% Change)	Soybeans *2018 Final* (1,000 Acres)	 *2019 June*[2] (% Change)	 *2019 August*[3] (% Change)
Illinois	11,000	0%	-3%	10,800	-5%	-7%
Indiana	5,350	3%	-5%	5,950	-11%	-9%
Kansas	5,450	8%	17%	4,750	-1%	-3%
Michigan	2,300	0%	-13%	2,300	-9%	-24%
Missouri	3,500	-3%	-7%	5,850	-9%	-13%
Ohio	3,500	-6%	-20%	5,000	-6%	-16%
South Dakota	5,300	-9%	-15%	5,650	-22%	-38%

References and Data–Accessed September 15, 2019

(1) Reference: June 28, 2019 USDA NASS Agricultural Statistics Board Notice

(2) Reference: American Farm Bureau Federation–Groundtruthing USDA's June Acreage Report

(3) Author calculations based on Corn Data and Soybean Data in NASS August 2019 *Crop Production*

Data solutions to improve early estimates

To inform acreage models NASS uses

- Survey data
- USDA Farm Service Agency (FSA) administrative data
- Remote sensing data (spectral reflectance)

and has recently started to investigate the use of

- Temperature and precipitation data (PRISM)
- Soil moisture (Crop-CASMA)
- Grain price basis data (GeoGrain)
- Crop rotation patterns based on NASS Cropland Data Layer (CDL)

Modeling acreage with remote sensing data

Remote sensing technology provides a variety of data to assess the status of the agriculture

Walker and Sigman (1984) developed a statistical model to predict planted acreage at the county level when survey and satellite data are both available

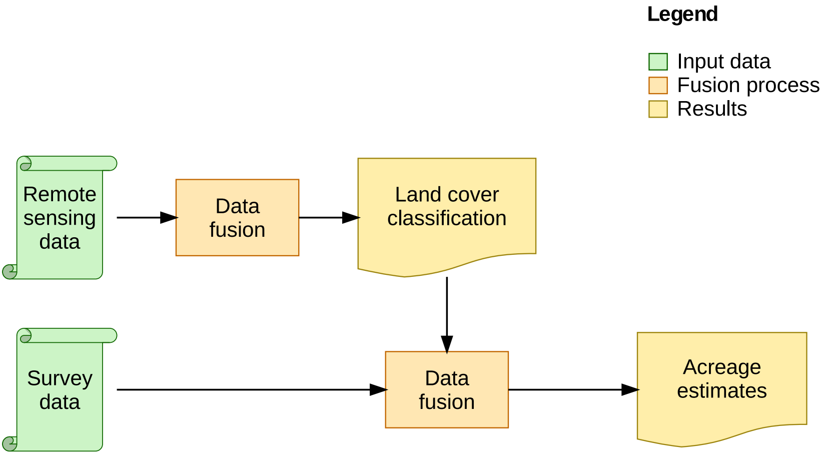
More recent challenges arise from

- Land-use and crop identification
- Non-parametric modeling
- Data processing practices

PART III

PROCESSING DATA SOURCES

A new way to provide early season estimates



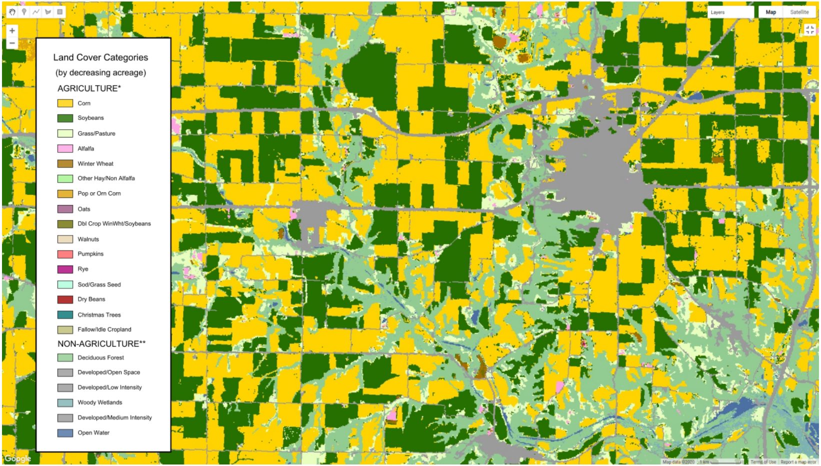
Remotely sensed images

Satellites provide data on visible light, near-infrared reflectance, surface temperature, soil moisture, and other radiometric measurements

- Data collection and spectrum encoding
  - Preprocessed by national and international space agencies
  - Further processing by private organizations
- Spatial resolution
  - Overlaying, snapping and re-projection
  - Masking and clipping
  - Sharpening
- Temporal resolution
  - Dependent on number of satellite acquisitions and orbits

Example of remotely sensed images (true colors)

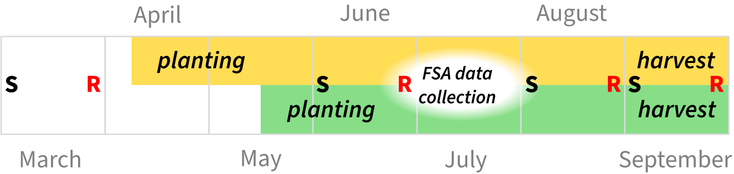
NASS CDL in December



Challenges of image fusion before June Acreage Report

- Using only images collected before July
- FSA training data
  1. Generally used for CDL production
  2. Not available in June for corn and soybeans

Legend
🟨 Corn
🟩 Soybeans
S Survey reference date
R Report due date

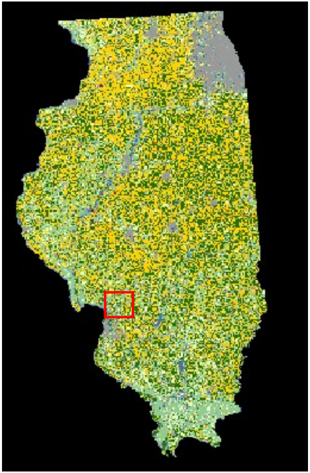


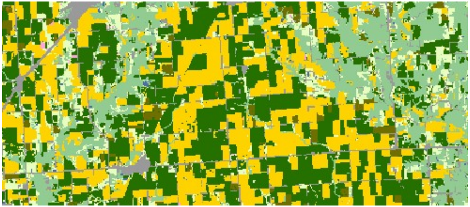
Early Season CDL (ESCDL)

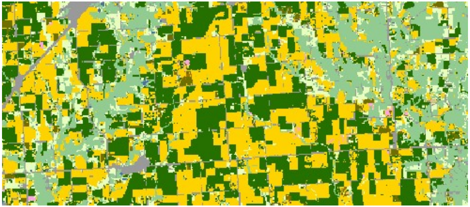
It has the potential to assess changes in planting intentions and drive the estimates towards more precise results

- Historical data are used for training
  - Input variables
    1. Crop rotation patterns (based the CDL)
    2. Remote sensing images from LANDSAT8 spectral bands within a time frame from March to mid-June
  - Output
    1. Crop classification
- Tree based models (Breiman et al., 1984) have been used to classify fields in Google Earth Engine (GEE)
- Validation performed with FSA ground reference data

Example of ESCDL produced in GEE


2018 June CDL - GEE


2018 June CDL - GEE


2018 – Final CDL

Fusing survey data and cropland classification results

The current fusion method is a small area approach developed by Walker and Sigman (1984)

$$\hat{Y}_s = N_s \left[\bar{y}_s + \hat{\beta}_s (\bar{X}_s - \bar{x}_s) \right]$$

$\hat{Y}_s$ total acreage estimates of stratum s

N_s number of units in stratum s

$\bar{y}_s$ sample mean from the survey data in stratum s

$\hat{\beta}_s$ robust regression adjustment

$\bar{X}_s$ population mean of stratum s from the CDL

$\bar{x}_s$ mean of stratum s from the CDL over sampled areas

The acreage estimates are obtained as the sum of estimated acreages over the strata

PART IV

CASE STUDY

Illinois pilot study

Illinois is a major corn and soybean producing state

The March Agricultural Survey provides acreage forecasts at the state level, which are based on surveyed farming intentions

In June, NASS conducts a survey and reports planted acreage estimates for each state in the nation

The Illinois pilot study is intended to identify the best statistical practice to provide early acreage estimates by combining survey and remote sensing data

ESCDL overall accuracies

ESCDL produced in 2017 by fusing data

- historical crop rotation patterns
- remotely sensed spectral reflectance

Accuracies are computed by comparing the ESCDL with FSA ground reference data

	Unfiltered	**Filtered**
ESCDL May	81.96	84.04
ESCDL June	82.88	84.80
NASS CDL	89.00	N/A

Estimate accuracy by fusing ESCDL and survey data

Estimates can be produced by regressing ESCDL field acreages with June Area Survey record level data (Walker and Sigman, 1984; Battese et al., 1988; Mueller and Seffrin, 2006).

Accuracies are computed as a relative difference with respect to the official NASS acreage estimates at the end of the year

Accuracy of IL planted acreage for corn

Year	June Area	Fusion
2016	-5.28 %	-1.80 %
2017	-4.17 %	0.08 %

Accuracy of IL planted acreage for soybeans

Year	June Area	Fusion
2016	-2.99 %	0.24 %
2017	-2.54 %	3.45 %

PART V

CONCLUDING REMARKS

Conclusion

- Identifying crop types at the field level is quite challenging before the crop has been planted and emerged
- The use of historical crop rotation patterns identified by the CDL has been beneficial in providing more accurate classification results at the field level
- Economic and extreme weather events have not been fully understood in the planting decision process, and the use of other data sources is part of this current research effort
- Preliminary analyses show the potential of data fusion in improving the accuracy of early season estimates by combining both survey and remote sensing data

Selected references

Battese, G. E. and Fuller, W. A. (1981). Prediction of county crop areas using survey and satellite data. In *Proceedings of the section on survey research methods, American Statistical Association*, volume 5, page 5. Am. Stat. Assoc Alexandria VA.

Battese, G. E., Harter, R. M., and Fuller, W. A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, 83(401):28–36.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and regression trees*. The Wadsworth statistics/probability series. Wadsworth & Brooks/Cole Advanced Books & Software, Monterey, CA.

Freund, Y., Schapire, R. E., et al. (1996). Experiments with a new boosting algorithm. In *icml*, volume 96, pages 148–156. Citeseer.

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., and Moore, R. (2017). Google earth engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*, 202:18–27.

Johnson, D. M. (2016). A comprehensive assessment of the correlations between field crop yields and commonly used modis products. *International Journal of Applied Earth Observation and Geoinformation*, 52:65–81.

Lenzerini, M. (2002). Data integration: A theoretical perspective. In *Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 233–246.

Mueller, R. and Seffrin, R. (2006). New methods and satellites: a program update on the nass cropland data layer acreage program. *Remote sensing support to crop yield forecast and area estimates, ISPRS archives*, 36(8):W48.

Walker, G. and Sigman, R. (1984). The use of landsat for county estimates of crop areas: evaluation of the huddleston-ray and the battese-fuller estimators for & the case of stratified sampling. *Communications in Statistics-Theory and Methods*, 13(23):2975–2996.

Zhang, J. (2010). Multi-source remote sensing data fusion: status and trends. *International Journal of Image and Data Fusion*, 1(1):5–24.

Zhao, J., Shi, K., and Wei, F. (2007). Research and application of remote sensing techniques in chinese agricultural statistics. Beijing. Paper presented at the Fourth International Conference on Agricultural Statistics.

Thank you!

Questions?

Luca Sartore, PhD	`lsartore@niss.org`
Jake Abernethy, PhD	`jake.abernethy@usda.gov`
Claire Boryan, PhD	`claire.boryan@usda.gov`
Lu Chen, PhD	`lchen@niss.org`
Kevin Hunt	`kevin.a.hunt@usda.gov`
Linda Young, PhD	`linda.j.young@usda.gov`