# Some Uses of R at USDA NASS

Nathan B. Cruze
nathan.cruze@nass.usda.gov

United States Department of Agriculture
National Agricultural Statistics Service (NASS)

EPA R Users Group Workshop
Washington, DC
September 11, 2017

# Me and My Agency

About me

- Mathematical statistician
- Research and Development Division, Sampling and Estimation Research Section
- Day-to-day applied research activities often begin with R

About USDA NASS

- Conducts Census of Agriculture every 5 years
- Conducts over 400 surveys annually
- Primarily a SAS shop

# Project I: Iterative Sequential Regression (ISR)

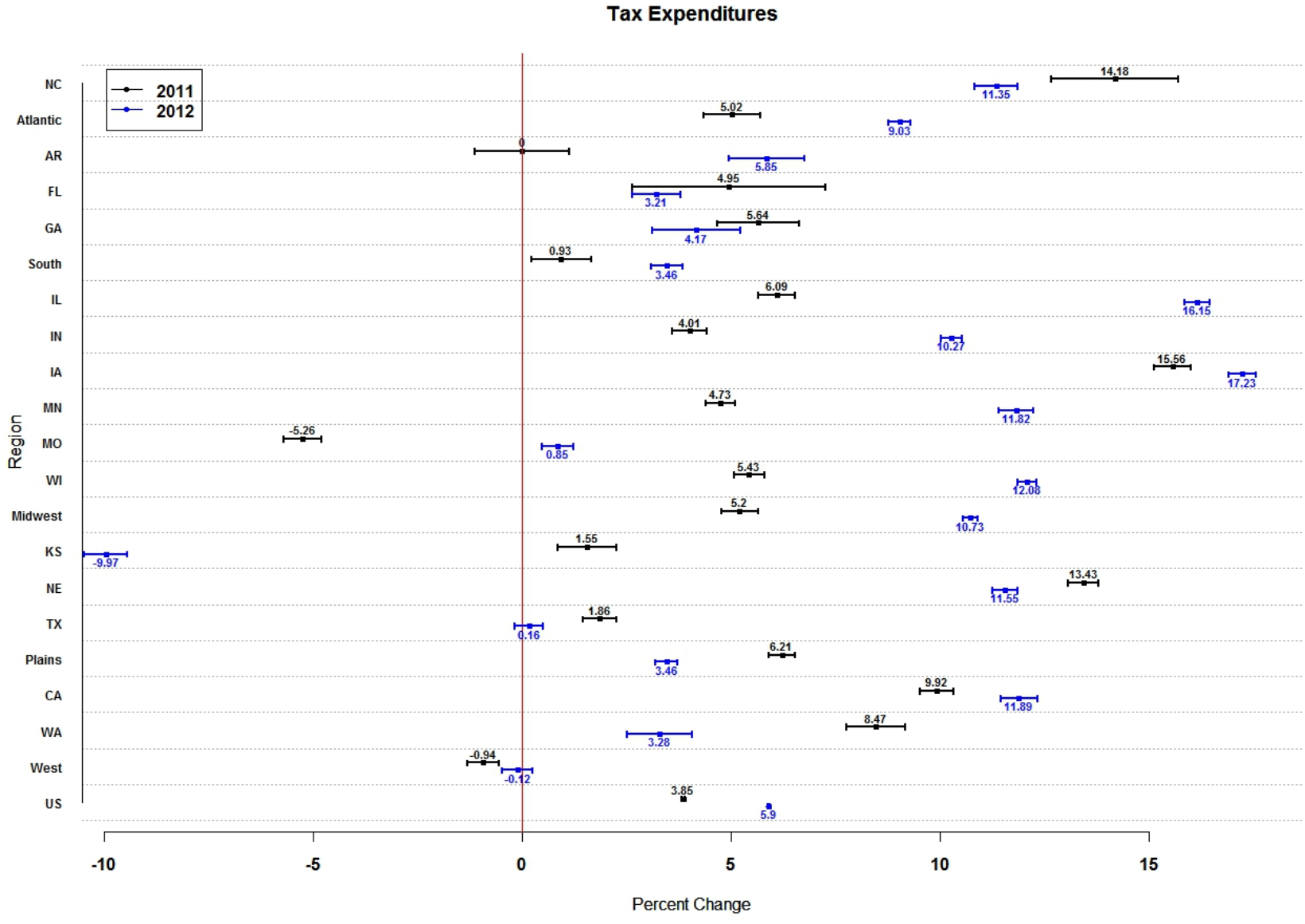
NASS conducts Agricultural Resource Management Survey (ARMS) in collaboration with USDA's Economic Research Service

**Main Idea:** Proposed ISR imputation methodology improves distributional characteristics and helps preserve relationships in the data

- Data augmentation and fully conditional specification
- Bayesian technique implemented through Gibbs sampling

Development involved multiple stakeholders and extended parallel testing

# Project I: Iterative Sequential Regression (ISR)



# Project I: Iterative Sequential Regression (ISR)

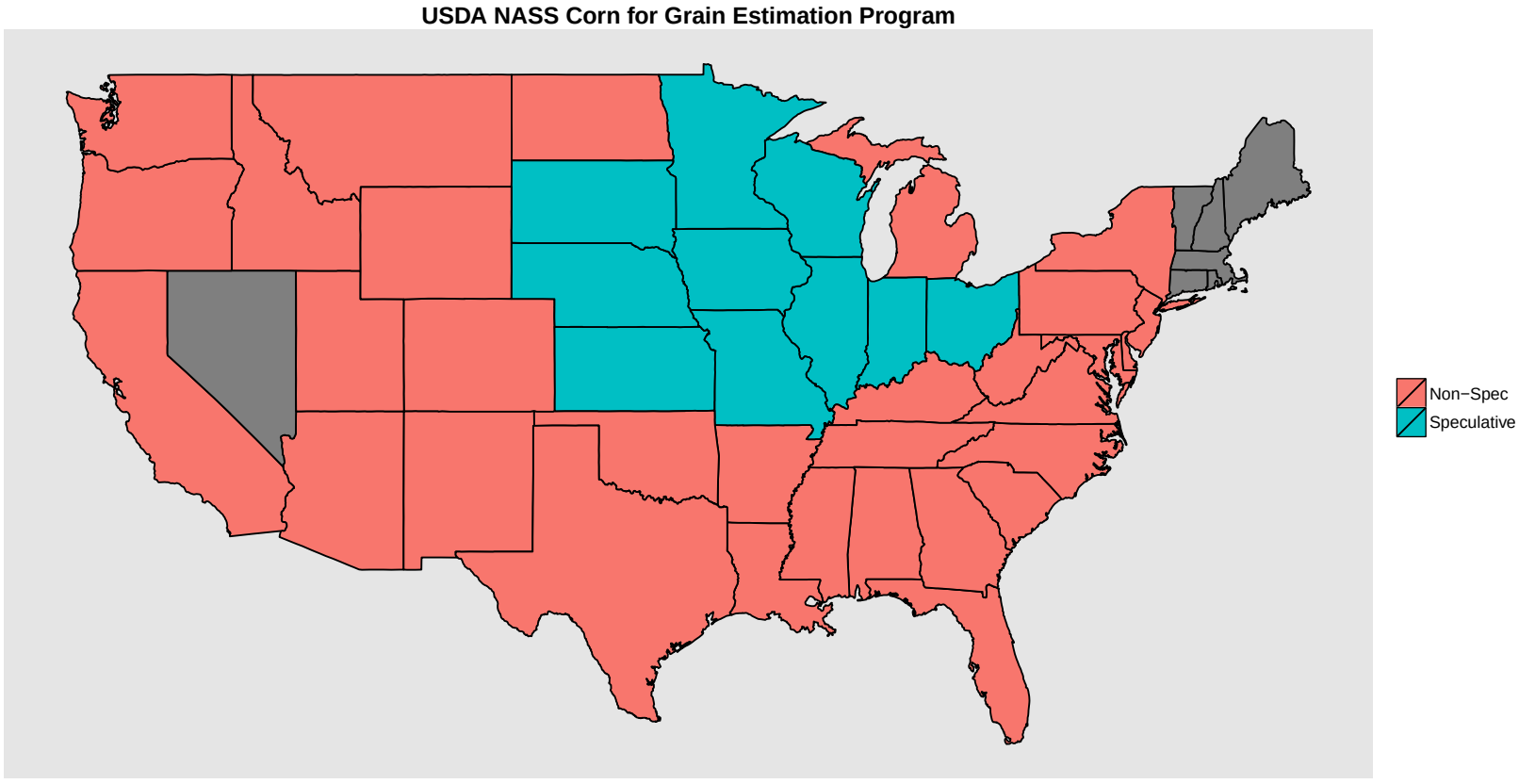
Development and testing resulted in the R scripts used in practice

Conditional mean imputation 'module' replaced with ISR module

- ISR machine imputation performed on a separate Linux box
  - Interaction with existing infrastructure
  - Written in R and C code
  - File input-output with SAS interface
- ISR output subsequently loaded for editing

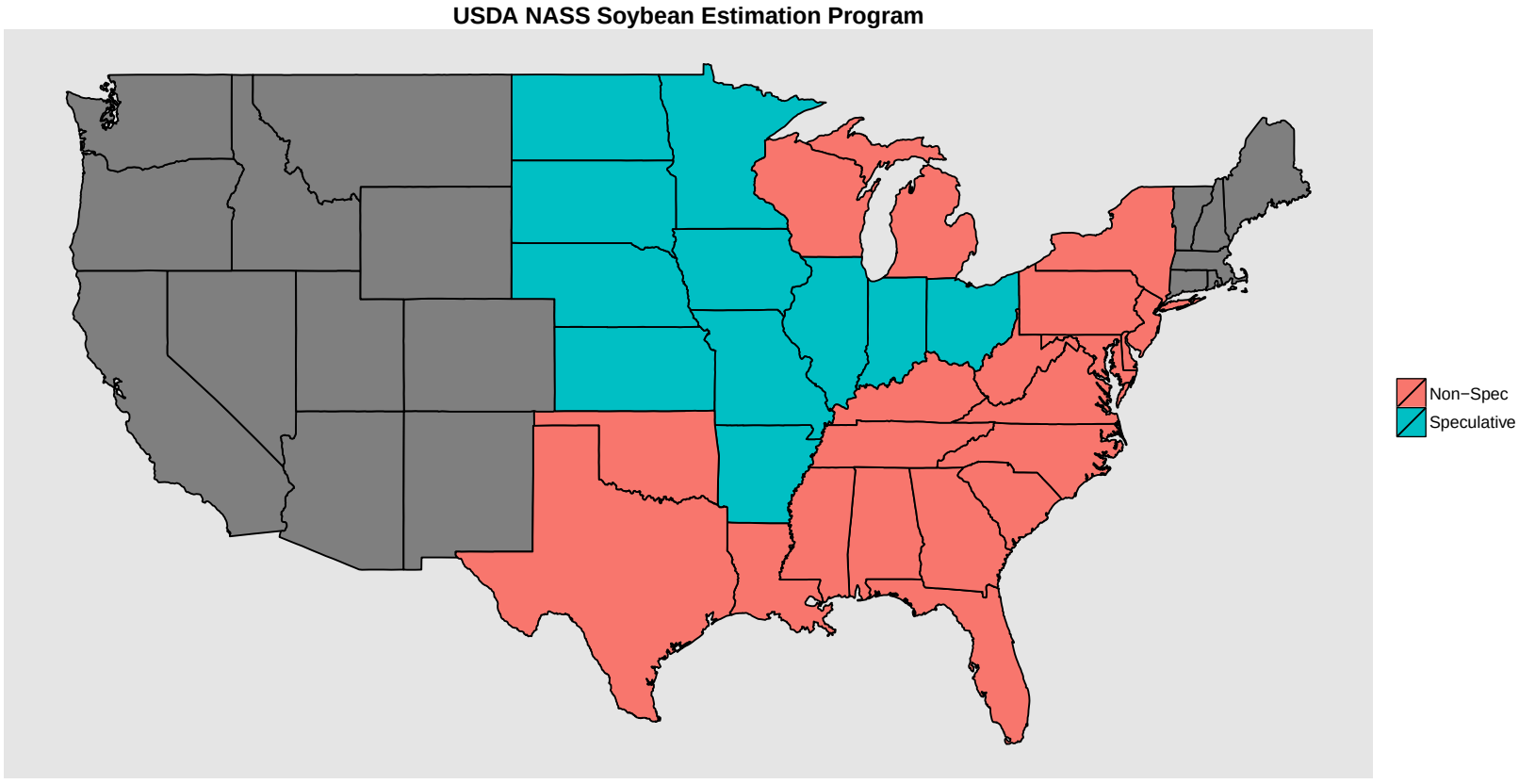
**R has been used to execute ISR imputation in production of ARMS Phase III estimates for the past three years**

# Basic Mapping: Corn for Grain



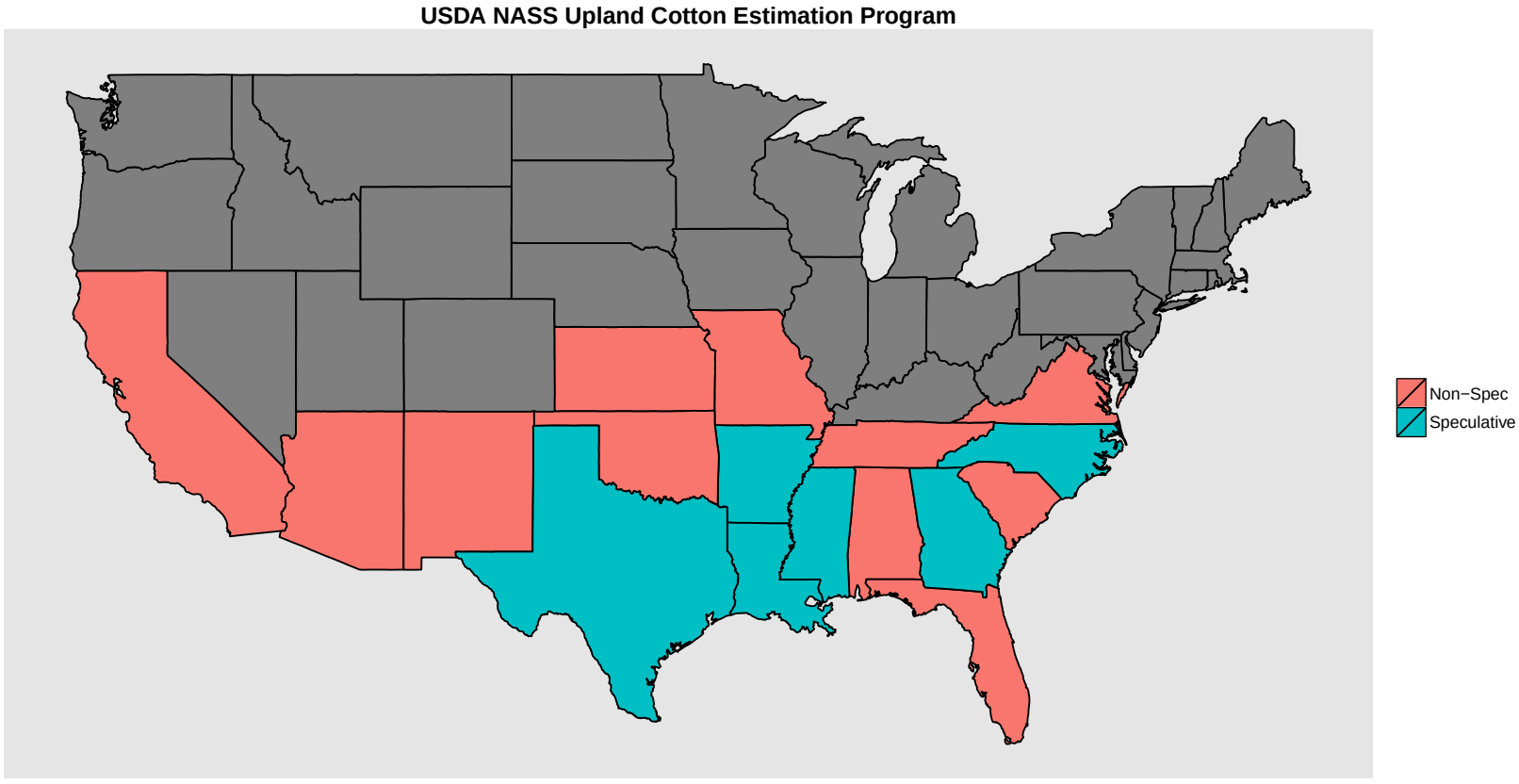

- Annual program: 41 states (production of grain)
- Speculative region: 10 states
- Yield forecasts: August-November, January (final)

# Basic Mapping: Soybeans



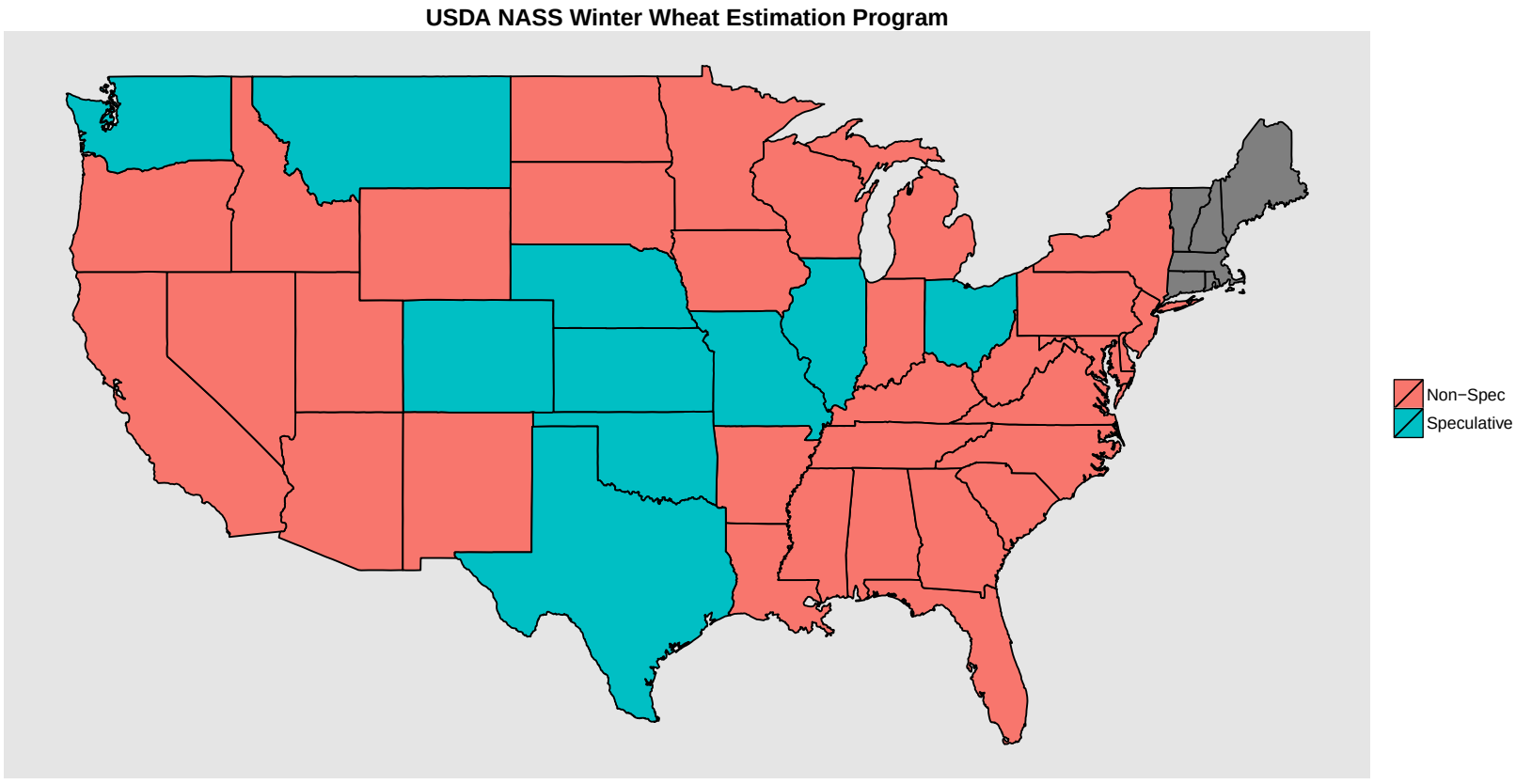

- Annual program: 31 states
- Speculative region: 11 states
- Yield forecasts: August-November, January (final)

# Basic Mapping: Upland Cotton



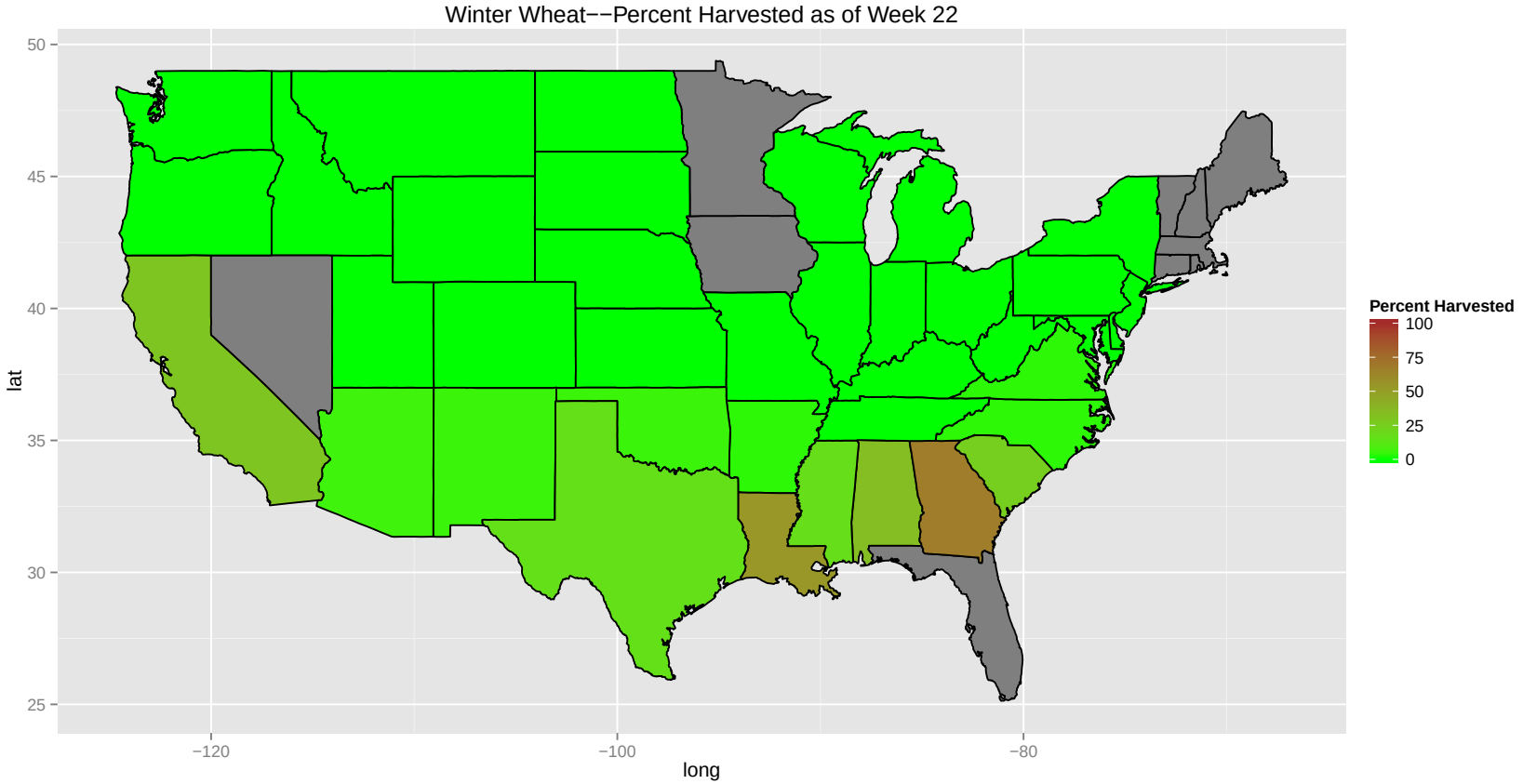

- Annual program: 17 states
- Speculative region: 6 states
- Yield forecasts: August-January, May (final)

# Basic Mapping: Winter Wheat

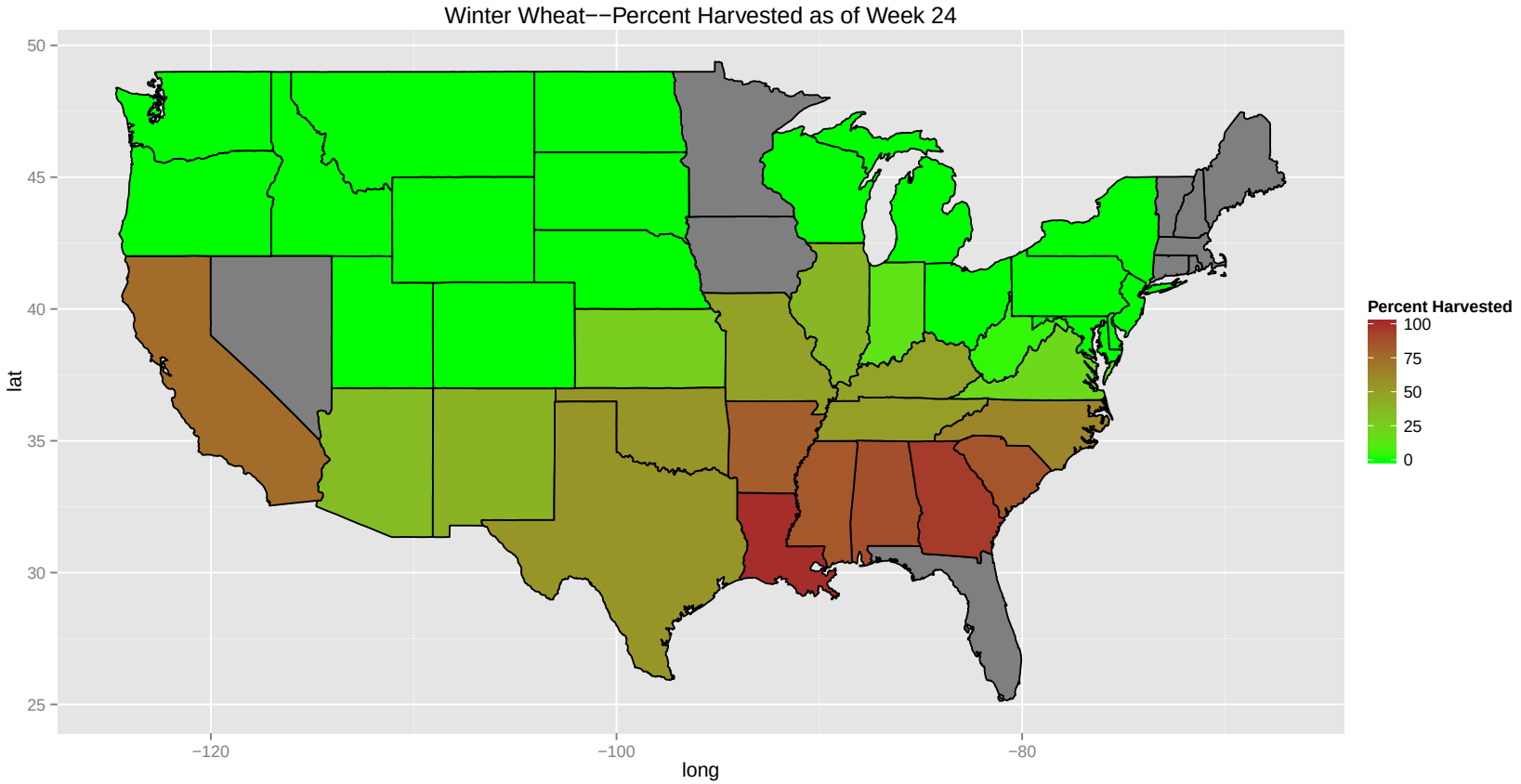



- Annual program: 42 states
- Speculative region: 10 states
- Yield forecasts: May-September (final)

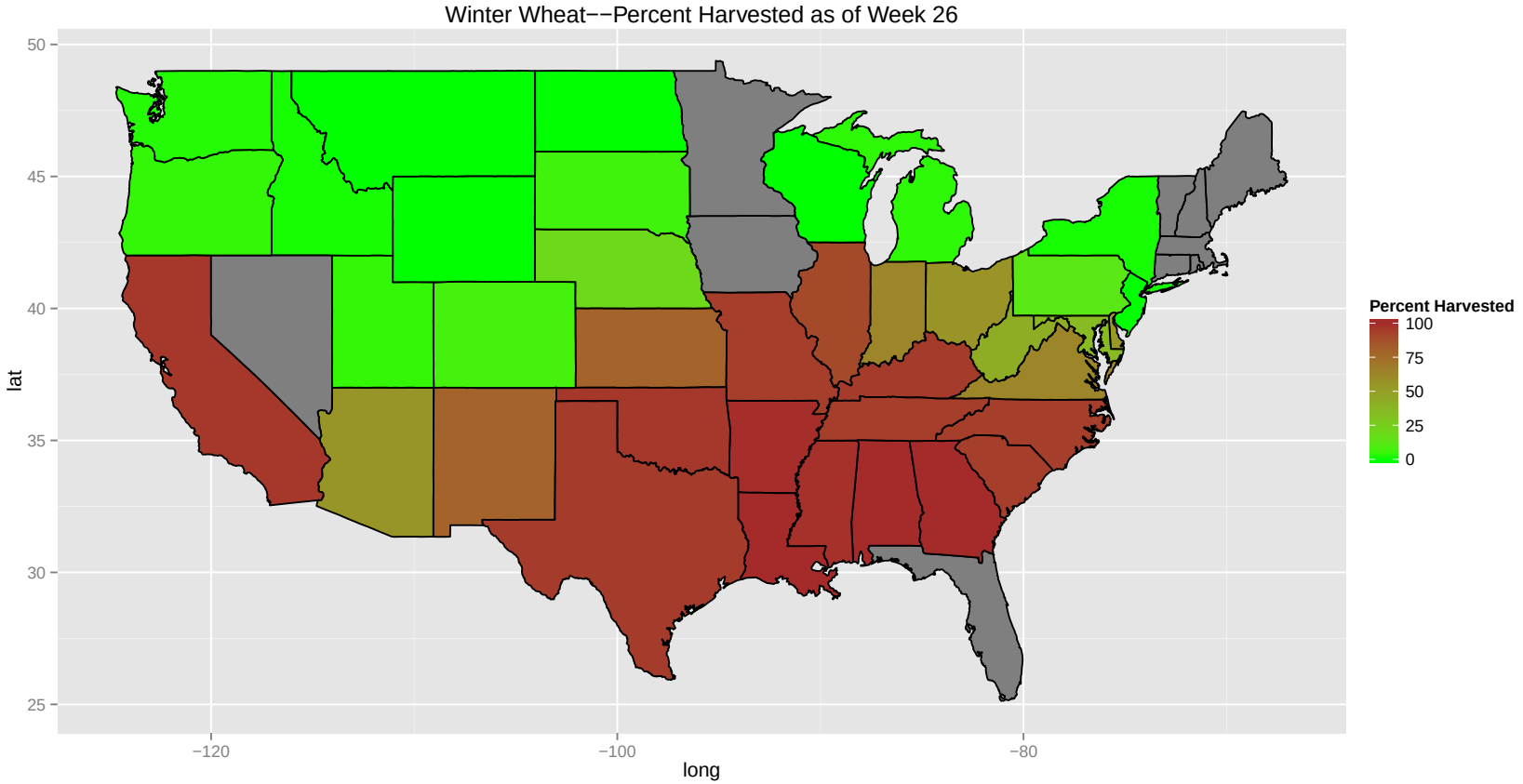
# Basic Mapping: Winter Wheat Harvest Progress



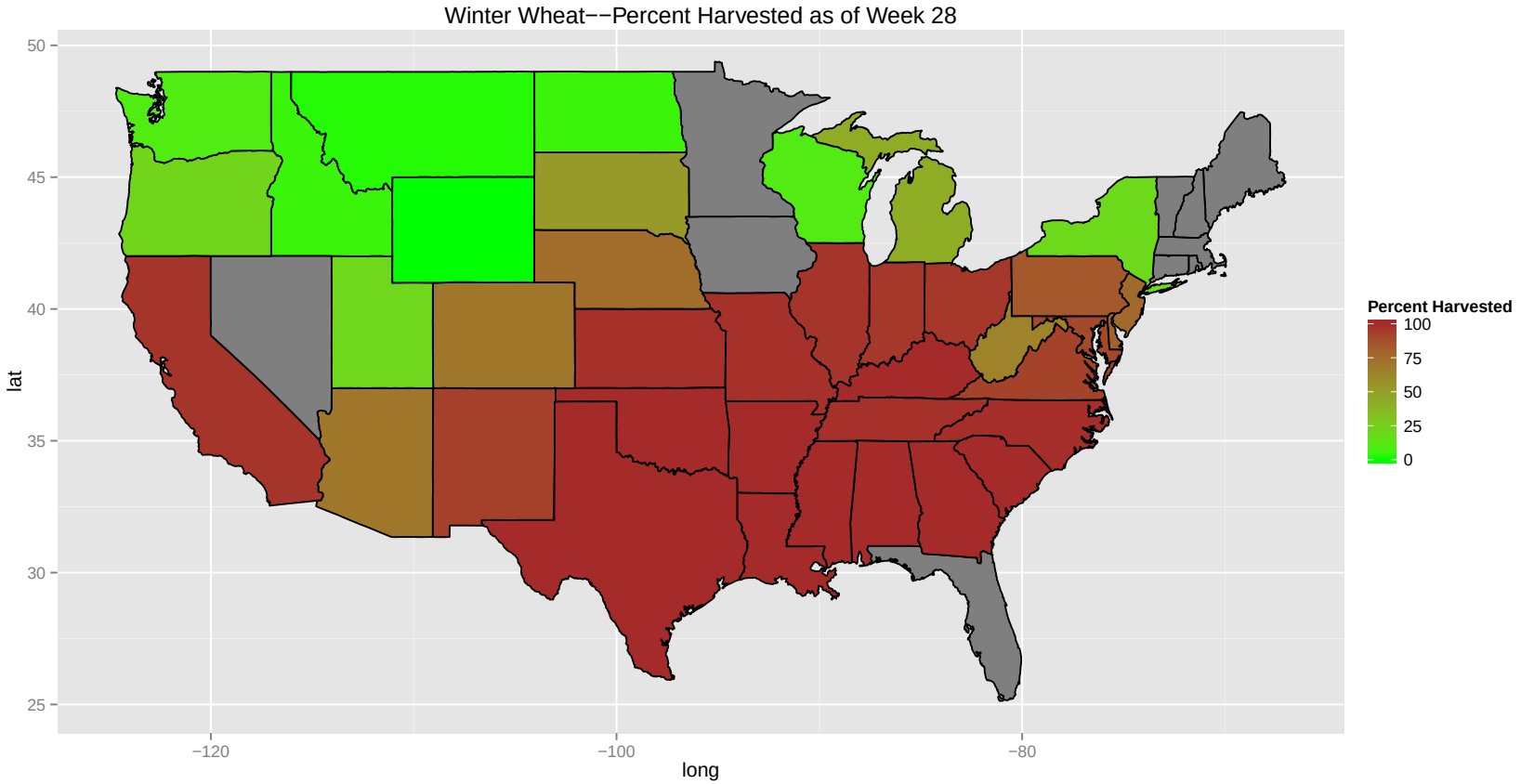
# Basic Mapping: Winter Wheat Harvest Progress



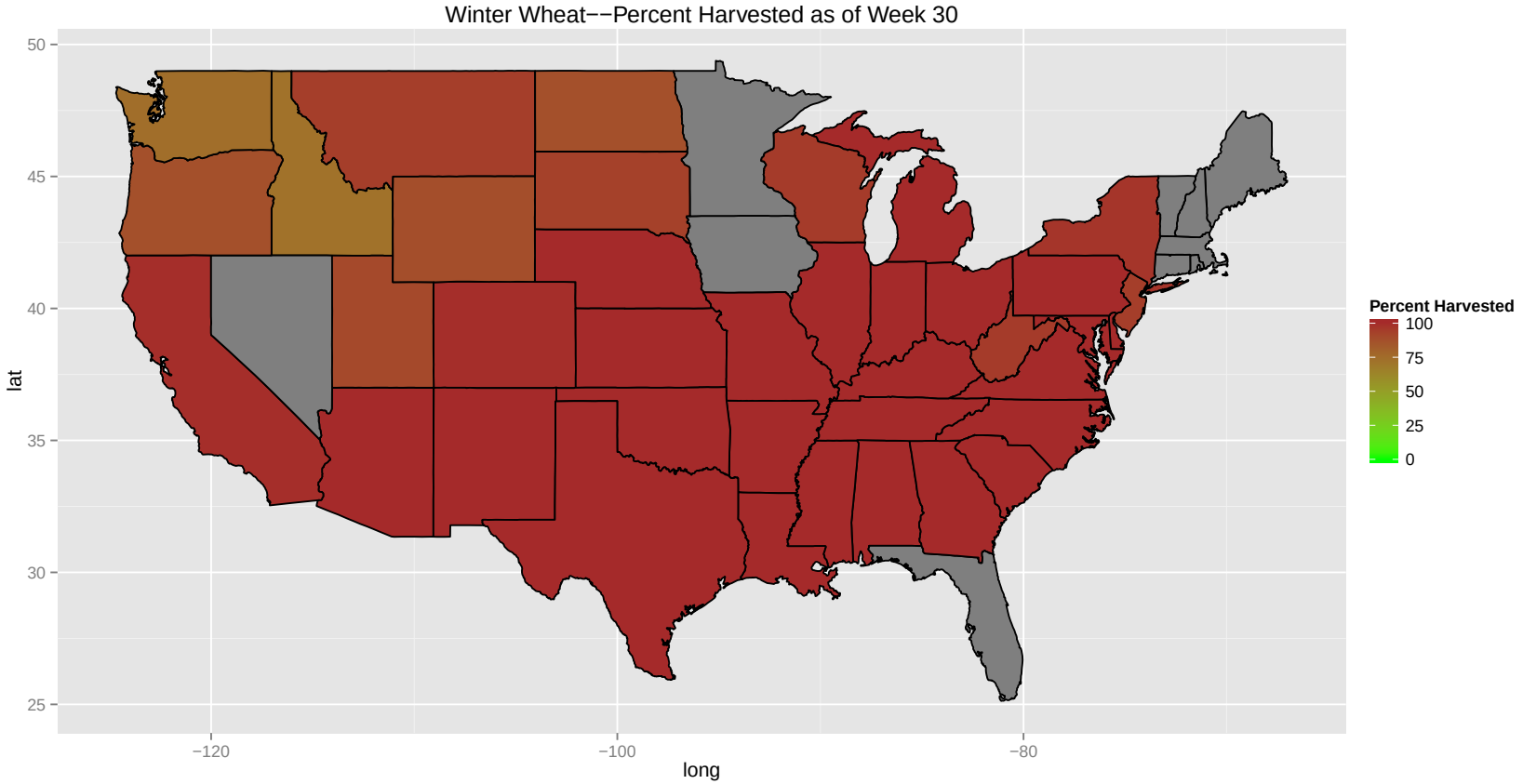
# Basic Mapping: Winter Wheat Harvest Progress



# Basic Mapping: Winter Wheat Harvest Progress



# Basic Mapping: Winter Wheat Harvest Progress

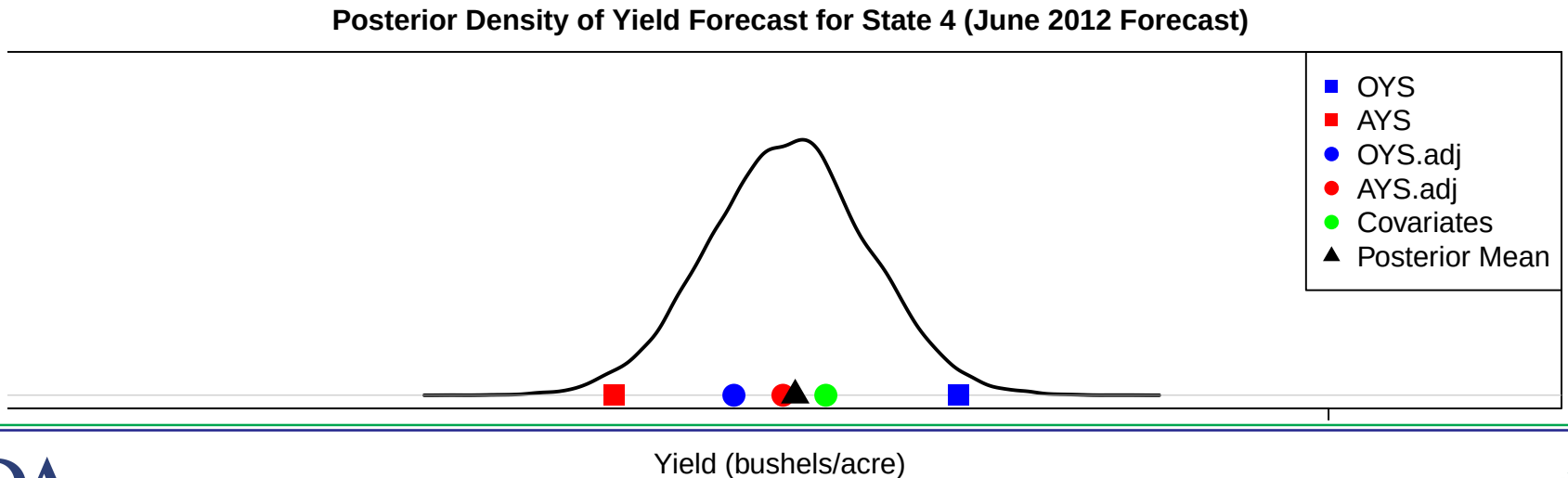


# Project II: Crop Yield Forecasting

Official NASS crop yield forecasts in Crop Production Report represent consensus of NASS Agricultural Statistics Board (ASB)

**Main Idea:** Develop Bayesian hierarchical models to produce one-number forecasts

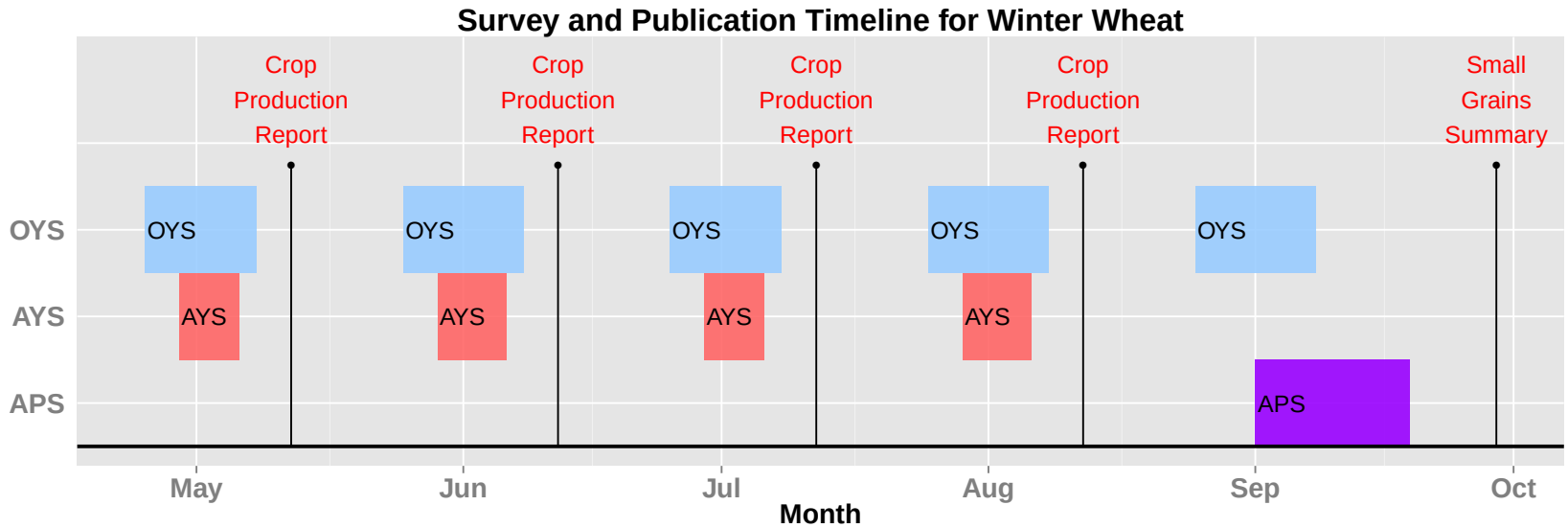
- Synthesize several survey estimates
- Produce measures of uncertainty
- Gibbs sampler used to obtain Monte Carlo estimates
- Currently we support speculative regions and member states



# Project II: Crop Yield Forecasting

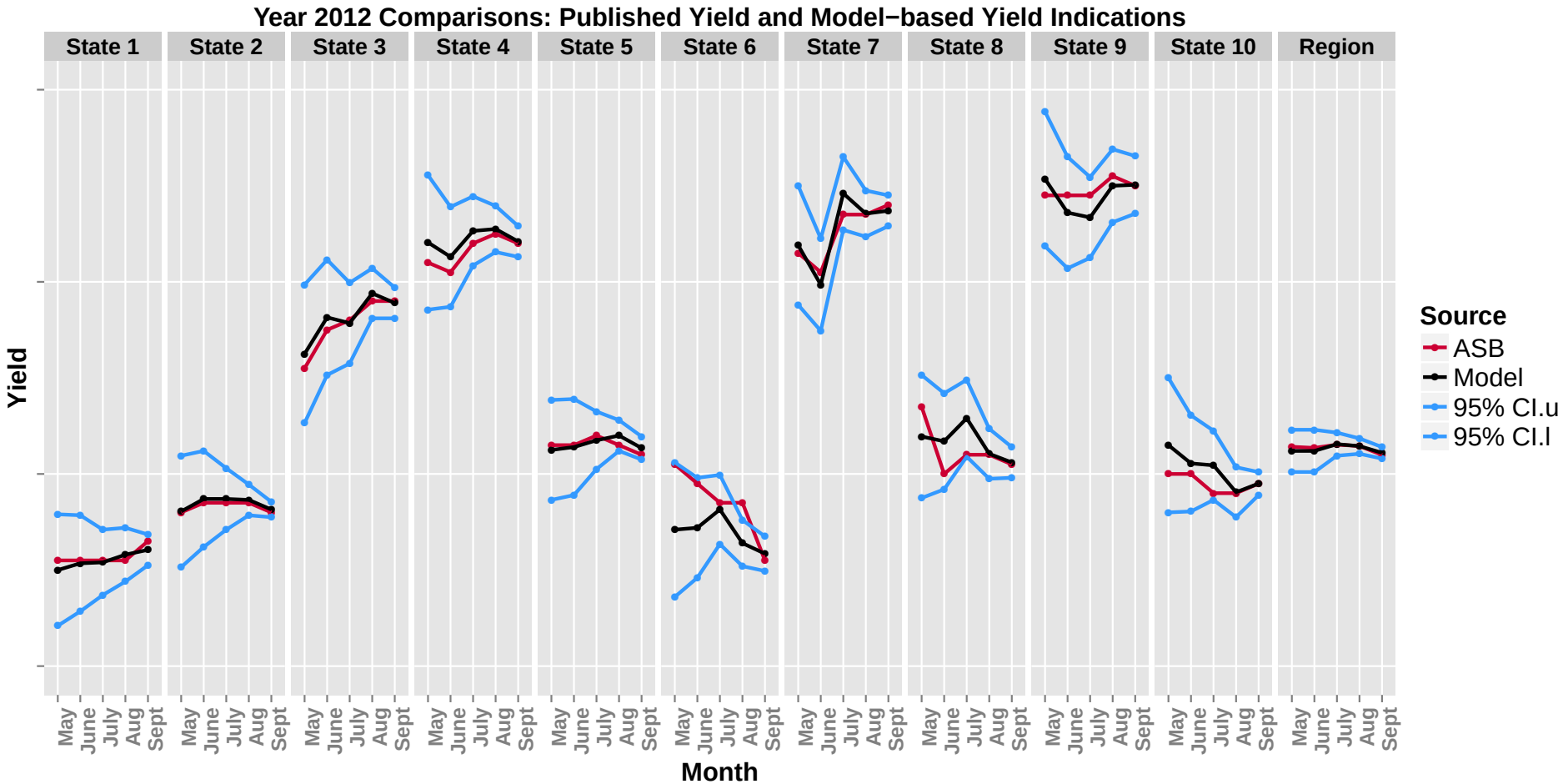
**Use in Production:** Research staff members have used R as a 'conduit' to execute a Gibbs sampler

- R calls C code for corn and soybean yield models (2011) and winter wheat yield models (2015)
- R calls JAGS for new upland cotton yield models (2017)
- **Modeled estimates are provided to Agricultural Statistics Board for their deliberations**



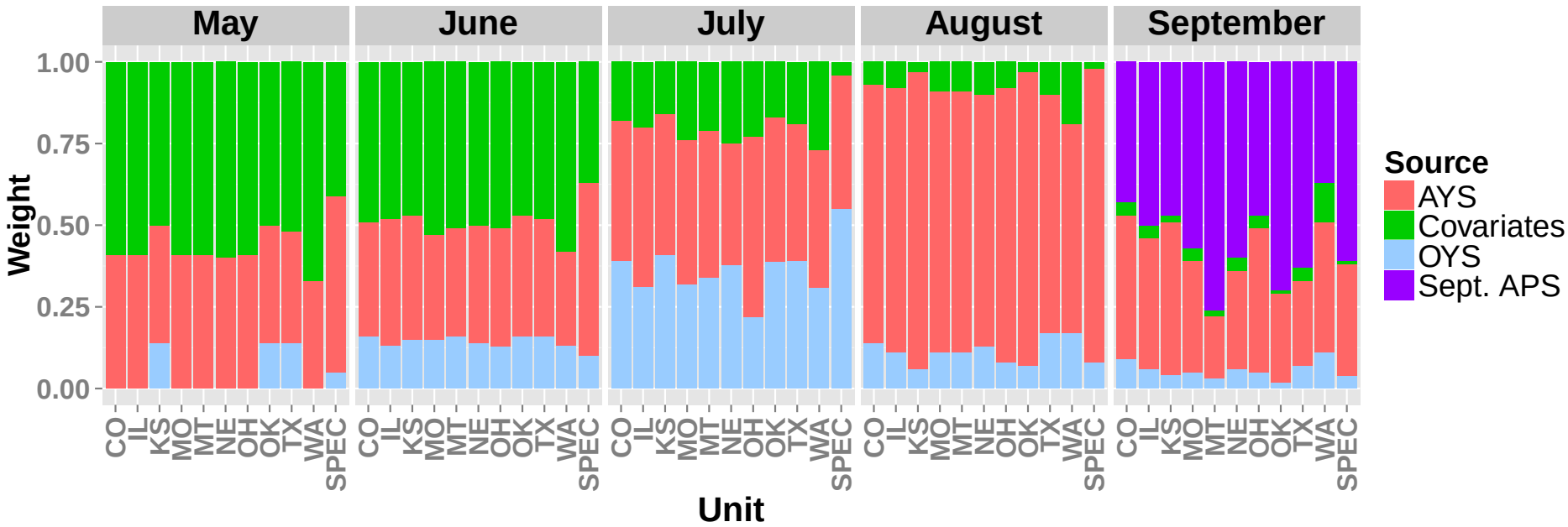
# Project II: Crop Yield Forecasting

Visualizing the sequence of model-based forecasts versus NASS official statistics



# Project II: Crop Yield Forecasting

Visualizing a model-based 'rule-of-thumb' for the combination of several disparate estimates



# Project III: Integer Calibration for Dual System Estimation

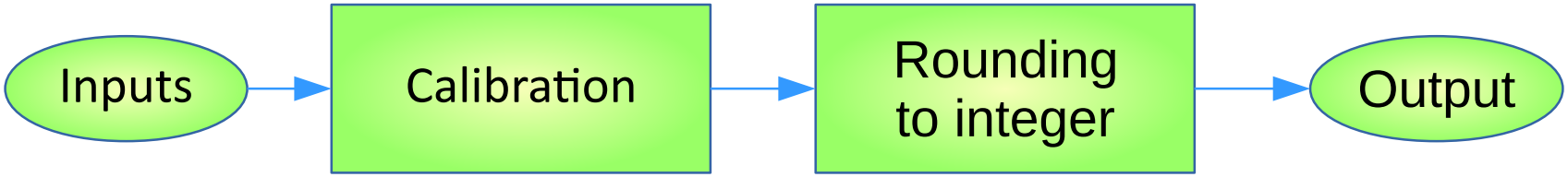
NASS adopted dual system estimation (DSE) approach in 2012 Census of Agriculture–applied DSE in other surveys

**Main Ideas:**

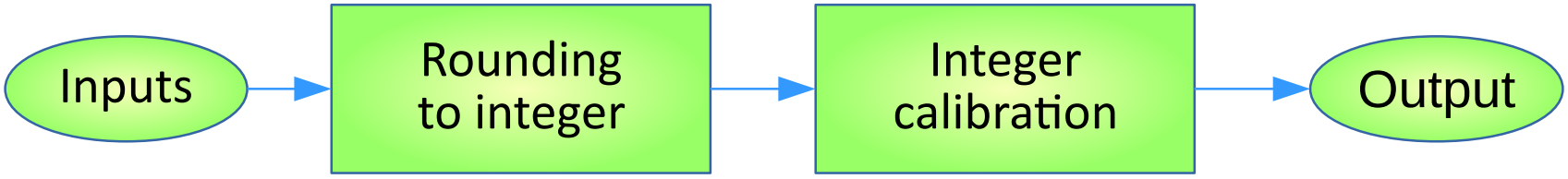
- DSE weights help adjust for undercoverage, nonresponse, misclassification
- Desire integer weights–consistent totals
- Desire estimates that satisfy many calibration targets
- Integer Calibration (INCA) improves rounding of weights and satisfaction of multiple targets

# Project III: Integer Calibration for Dual System Estimation

Previous approach:



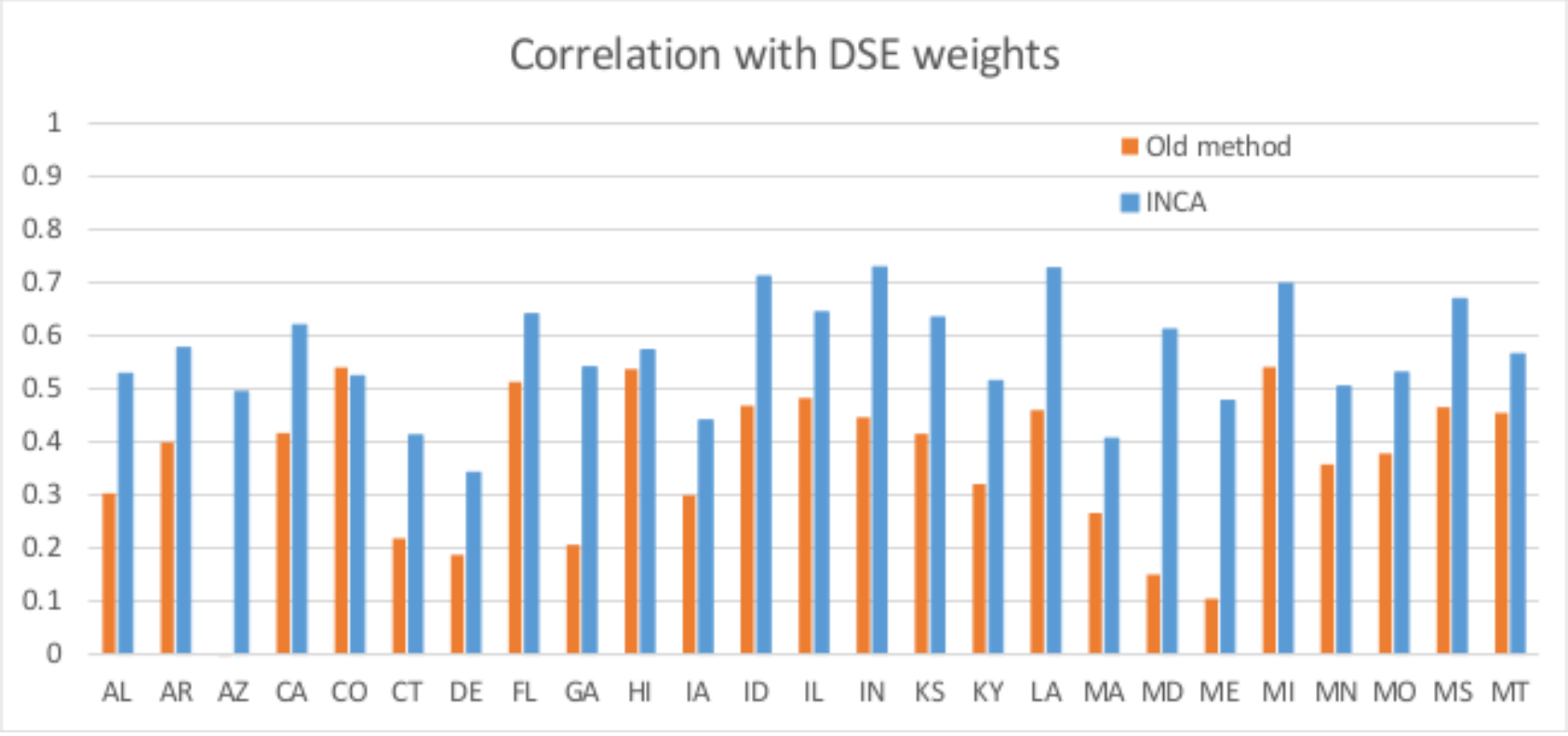

INCA:




**INCA was developed in R and the 'INCA' package exists on CRAN. DSE with INCA was executed in R for the the 2015 Local Foods Survey.**

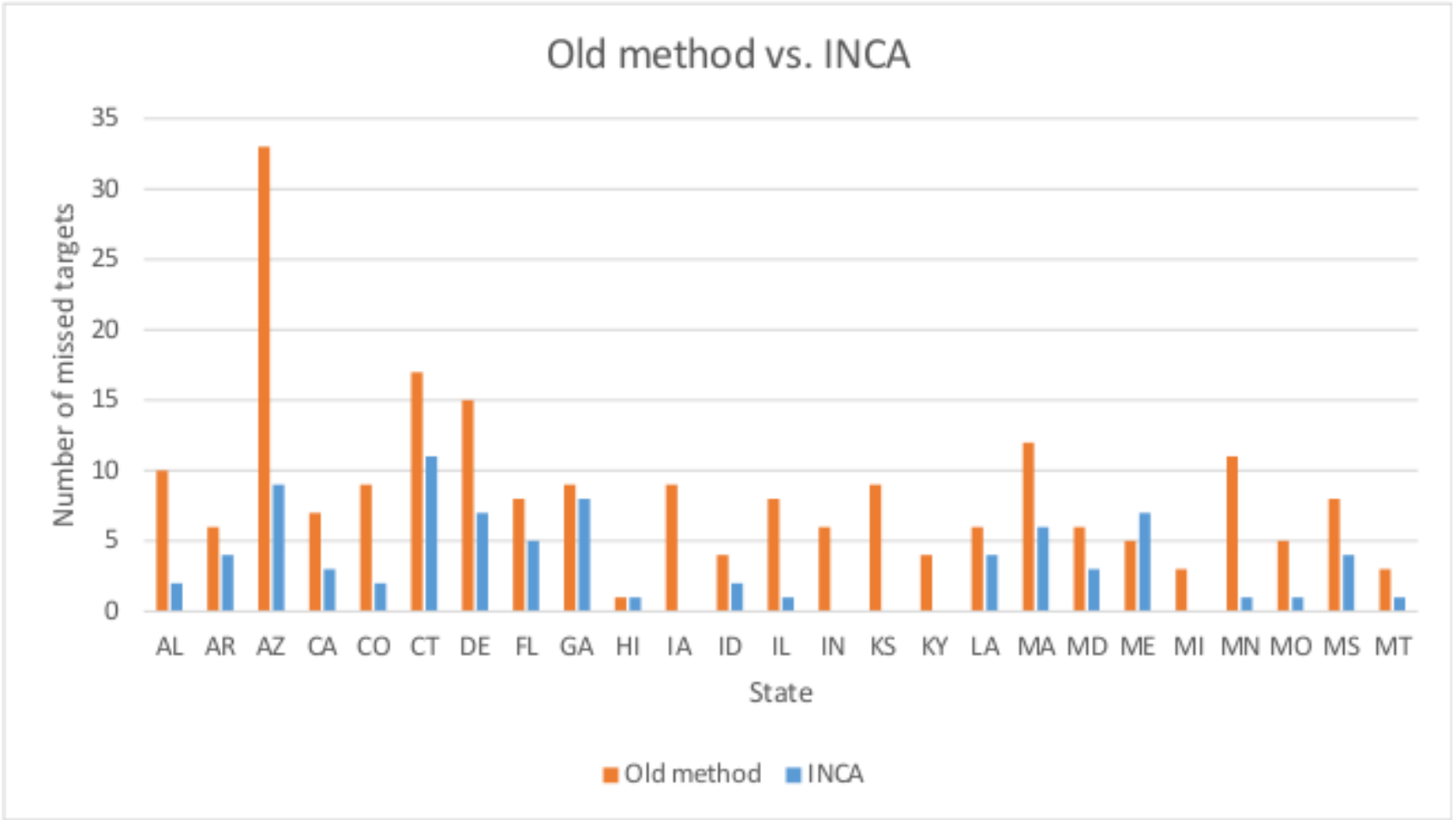
# Project III: Integer Calibration for Dual System Estimation

INCA weights are 'closer' to (more highly correlated with) original DSE weights



# Project III: Integer Calibration for Dual System Estimation

INCA weights ensure that estimated totals miss fewer calibration targets



# Additional Production and Research Projects in R

R in Production

1. Small area models for cash rental rates (2013)
2. Quarterly model-based estimates of cattle inventory
3. Simulated annealing for substratification of primary sampling units in June Area Survey

R in Research and Development

1. Quarterly model-based estimates of hog inventory
2. Small area crop acreage, yield, and production estimates

# References


Cruze, N. B. (2016). A Bayesian hierarchical model for combining several crop yield indications. In JSM Proceedings, Survey Research Methods Section. Alexandria, VA: American Statistical Association.

Miller, D., Dau, A., and Lisic, J. (2015). Comparison of modern imputation methodologies on complex data from agricultural operations. In Proceedings of the 2015 FCSM Research Conference [online, accessed 10-sept-2017]. https://www.nass.usda.gov/Education_and_Outreach/Reports,_Presentations_and_Conferences/reports/conferences/FCSM/D1_Miller_2015FCSM.pdf.

Nandram, B., Berg, E., and Barboza, W. (2014). A hierarchical Bayesian model for forecasting state-level corn yield. *Environmental and Ecological Statistics*, 21(3):507–530.

Sartore, L., Toppin, K., and Spiegelman, C. (2016). Integer programming for calibration. In Presentations of 2016 Federal CASIC Workshops [online, accessed 10-sept-2017]. https://www.census.gov/fedcasic/fc2016/ppt/1_5_Integer.pdf.

Wang, J. C., Holan, S. H., Nandram, B., Barboza, W., Toto, C., and Anderson, E. (2012). A Bayesian approach to estimating agricultural yield based on multiple repeated surveys. *Journal of Agricultural, Biological, and Environmental Statistics*, 17(1):84–106.