On Combining Multiple Sources of Information to Enhance NASS Crop Estimates

Nathan B. Cruze

National Agricultural Statistics Service (NASS)
United States Department of Agriculture
nathan.cruze@nass.usda.gov

CNSTAT Panel on Improving Federal Statistics for Policy and Social Science Research Using Multiple Data Sources and State-of-the-Art Estimation Methods

December 16, 2015

Presentation Outline

1. NASS crop survey cycle–winter wheat
2. The Agricultural Statistics Board–synthesizing many sources of information
3. Model-based forecasts of winter wheat yield
4. Recurring themes and challenges of combining multiple sources of information

Winter Wheat Production Cycle

Month	Crop Stage	NASS Surveys	NASS Publications	Variables of Interest
December	Most planting complete	December APS	Winter Wheat Seedings Grain Stocks	Acres Planted On & Off Farm Stocks
March	Emerging from dormancy	March APS	Prospective Plantings Grain Stocks	Acres Planted On & Off Farm Stocks
May	Crop actively developing	May AYS May OYS	Crop Production	Acres Harvested, Yield, Production
June	Crop develops; early harvest	June AYS June OYS June APS June Area Survey	Crop Production Acreage Grain Stocks	Acres Harvested, Yield, Production Acres Planted/Harvested Ending On & Off Farm Stocks
July	Harvest underway	July AYS July OYS	Crop Production	Acres Harvested, Yield, Production
August	Harvest winding down	August AYS August OYS	Crop Production	Acres Harvested, Yield, Production
September	Harvest complete	September APS	Small Grains Summary Grain Stocks	Acres Planted/Harvested, Yield, Production Beginning On & Off Farm Stocks
August-December	Marketing activities	Small Grains CAPS	County Estimates	Acres Planted/Harvested, Yield, Production

- Sequence of survey indications–update our understanding
- Multiple survey indications in same month
- Administrative data, remote sensing indications, weather data
- **One official statistic in publication**

The Agricultural Statistics Board (ASB)

Agricultural Statistics Board–panel of NASS commodity specialists

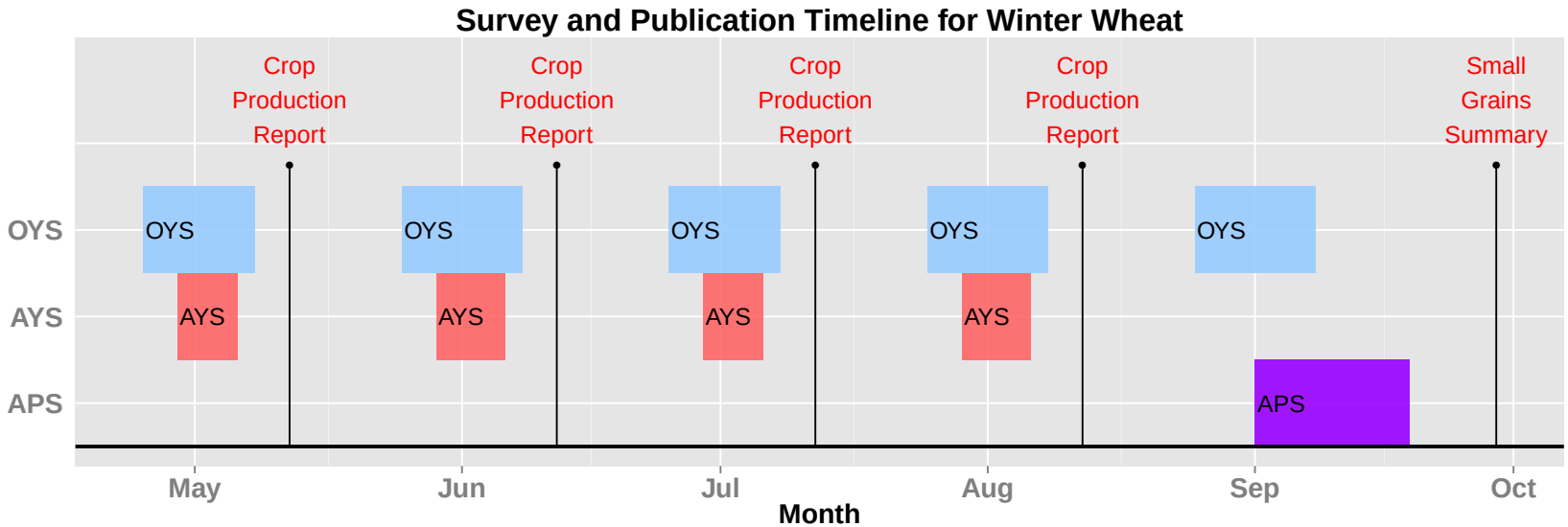
- Review current and historical survey 'indications'
- Review other available information–weather data, crop condition indications, administrative data
- Consensus on estimates of parameters of interest–planted/harvested area, production, yield
- **Official statistics in Crop Production Report are a composite or synthesis of multiple sources of information determined by expert assessment**

Challenge: Capture expert assessment in a manner that is
1) easily reproducible
2) includes appropriate measures of uncertainty

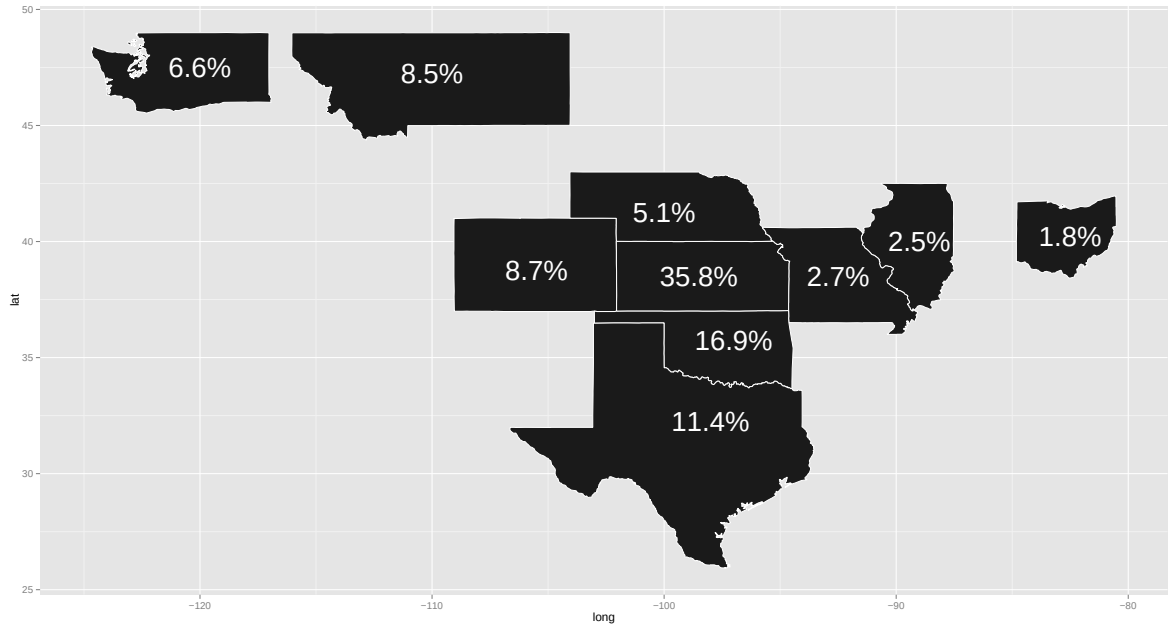
A Bayesian Hierarchical Model for Crop Yield Forecasting

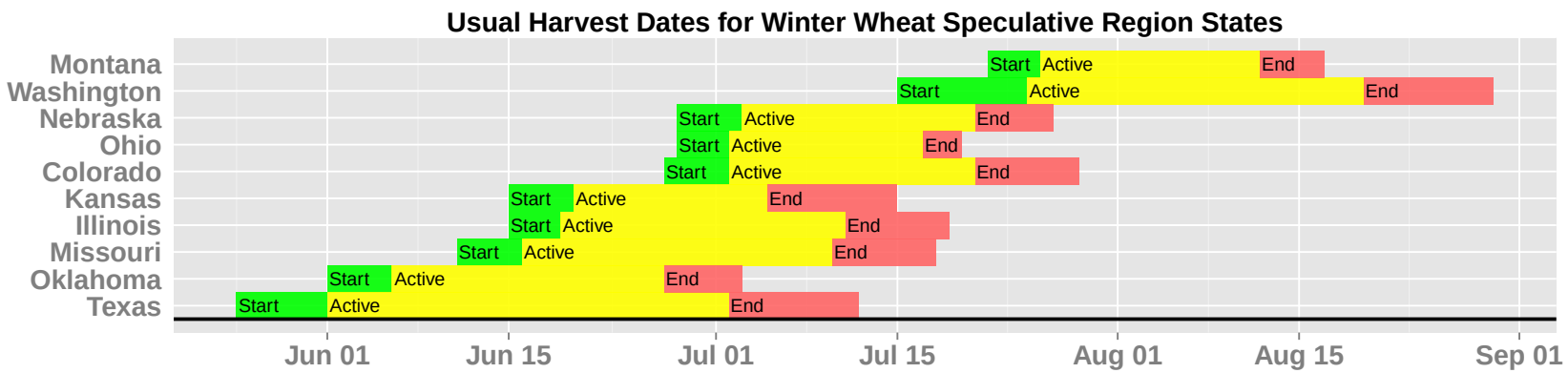
Goal: Synthesis of yield indications from NASS surveys and auxiliary information

- Yield measures output per area harvested (bushels/acre)
- Yield for state j: μ_j, $j = 1, 2, ..., J$
- Yield for **speculative region**: $\mu = \sum_{j=1}^{J} w_j \mu_j$
- Weights $w_j \propto$ harvested acres for state j

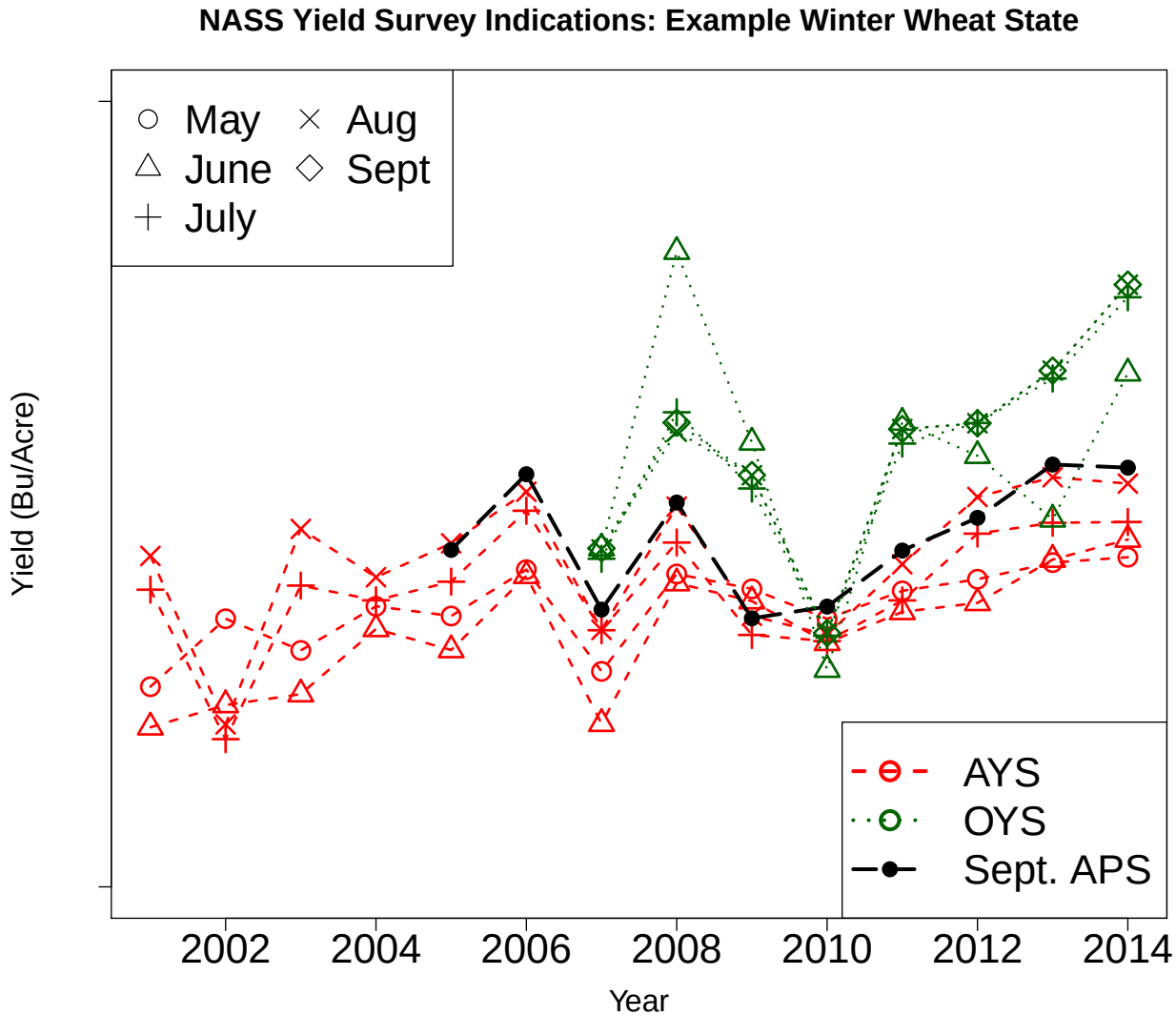


Winter Wheat Speculative Region–Acreage and Dates





Winter Wheat Yield Data–Example State



Bayesian Hierarchical Model for Speculative Region

Notation

- μ_t–true yield
- y_{ktm}–observed yield
- $k \in \{O, A, Q\}$–survey index
- $t \in \{1, ..., T\}$–year index
- $m \in \{months\}$–survey month
- $m*$–forecast month

Region data model

$$y_{ktm*}|\mu_t \sim \text{indep } N\left(\mu_t + b_{km*}, s^2_{ktm*} + \sigma^2_{km*}\right), k = O, A \quad (1)$$

$$y_{Qt}|\mu_t \sim \text{indep } N\left(\mu_t, s^2_{Qt}\right) \quad (2)$$

Region process model

$$\mu_t \sim \text{indep } N\left(\boldsymbol{z}_t'\boldsymbol{\beta}, \sigma^2_\eta\right) \quad (3)$$

Diffuse prior distributions

- Data model parameters: $\boldsymbol{\Theta}_d \equiv \left(b_{km*}, \sigma^2_{km*}\right)$
- Process model parameters: $\boldsymbol{\Theta}_p \equiv \left(\boldsymbol{\beta}, \sigma^2_\eta\right)$

Bayesian Hierarchical Model–Speculative Region Yield

Likelihood function–assuming conditional independence

$$[y_O, y_A, y_Q|\mu_t, \Theta_d] = \prod_{k\in\{O,A,Q\}} [y_k|\mu_t, \Theta_d] \tag{4}$$

Posterior distribution

$$[\mu_t, \Theta_d, \Theta_p|y_O, y_A, y_Q] \propto \prod_{k\in\{O,A,Q\}} [y_k|\mu_t, \Theta_d][\mu|\Theta_p][\Theta_d][\Theta_p] \tag{5}$$

Full conditional of regional yield, μ_t

$$[\mu_t|y_O, y_A, y_Q, \Theta_d, \Theta_p] \sim N\left(\frac{\Delta_2}{\Delta_1}, \frac{1}{\Delta_1}\right) \tag{6}$$

$$\Delta_1 = \sum_{k=O,A} \frac{1}{\sigma^2_{km*} + s^2_{ktm*}} + \frac{I_{\{Q\}}}{s^2_{Qt}} + \frac{1}{\sigma^2_\eta} \tag{7}$$

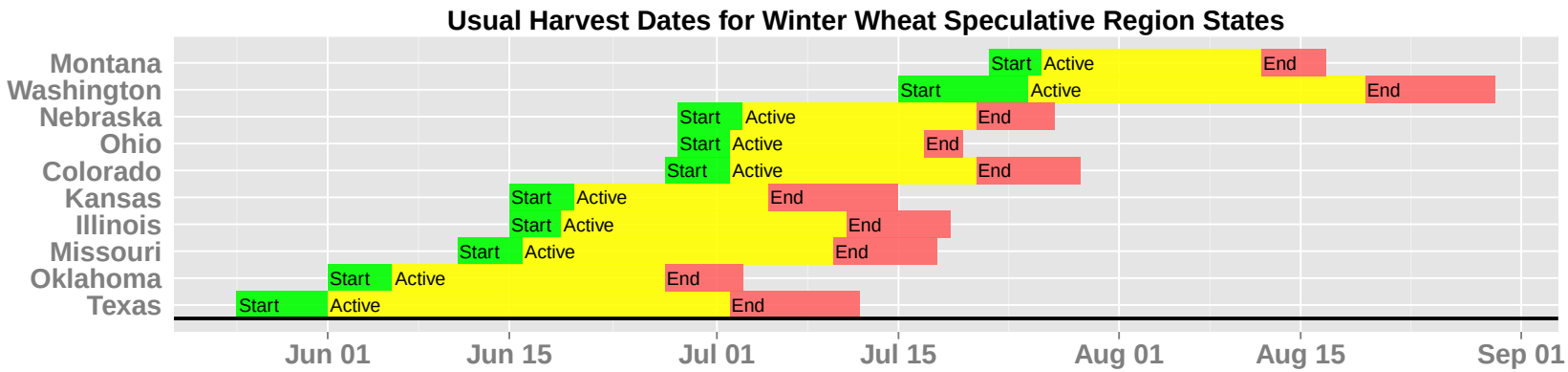
$$\Delta_2 = \sum_{k=O,A} \frac{y_{ktm*} - b_{km*}}{\sigma^2_{km*} + s^2_{ktm*}} + \frac{I_{\{Q\}} y_{Qt}}{s^2_{Qt}} + \frac{z_t'\beta}{\sigma^2_\eta} \tag{8}$$

Auxiliary Information

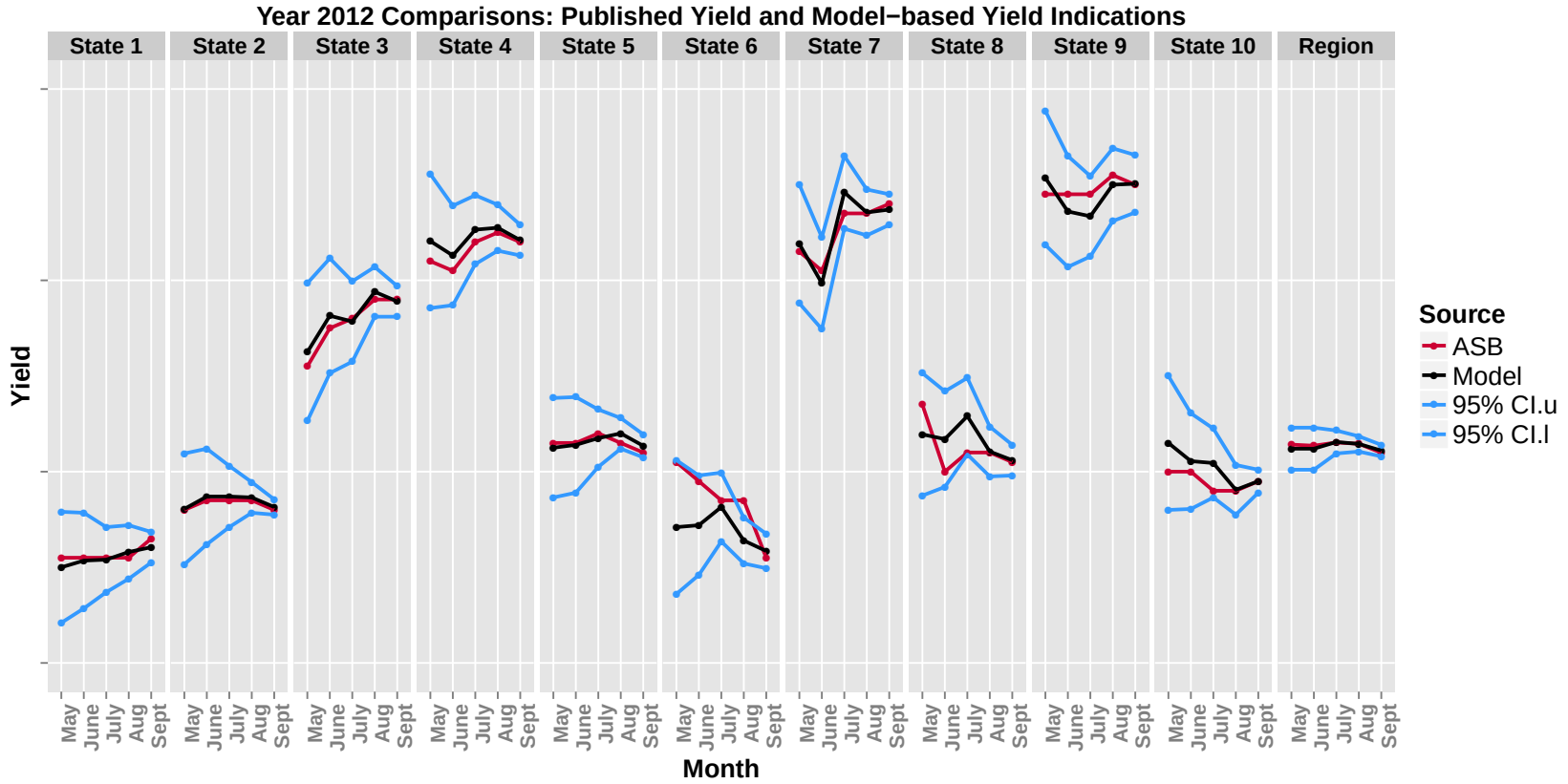
Covariates reflect conditions approaching active harvest dates

$$\mu_{tj} = \beta_{j1} + \beta_{j2} z_{j2} + \beta_{j3} z_{j3} + \beta_{j5} z_{j4} + \beta_{j5} z_{j5}$$

- State-specific constant
- z_{j2}: Linear time trend
- z_{j3}: Monthly precipitation (NOAA)
- z_{j4}: Monthly avg. temperature (NOAA)
- z_{j5}: Crop condition–% good + % excellent week # (NASS)



Comparing ASB Estimates and Model Outputs

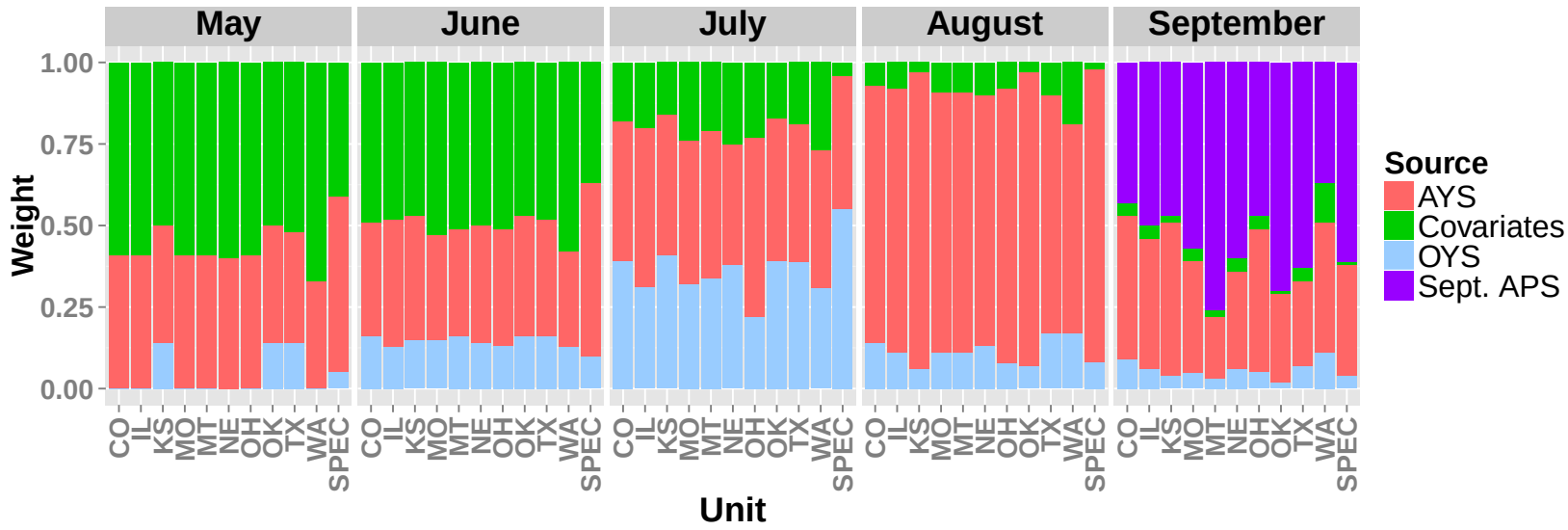


Emphasis Placed on Each Information Source

Weighted average decomposition of forecasts by information source

$$\hat{\mu}_{tj} \approx \sum_{k \in \{O,A,Q,Covariates\}} c_k (SOURCE)_k \qquad (9)$$

$$c_k \propto (variance)_k^{-1}$$



Many Data Sources–Recurring Themes and Challenges

1. NASS tradition–official statistic derived from many data sources
2. Pertains to all spatial scales, as well as livestock and economic programs
3. NASS is convening an expert panel through CNSTAT to strengthen crop county estimates and cash rental rates

Challenges

- Timeliness and availability of auxiliary information by areal unit, commodity, *variables of interest*
- Coverage/accuracy of administrative or remote sensing data
- Benchmarking and variance estimation
- Disclosure
- Tight publication deadlines–simplicity preferred!
- *Impact on end users*

Select References

Adrian, D. (2012). A model-based approach to forecasting corn and soybean yields. Fourth International Conference on Establishment Surveys.

Cruze, N. B. (2015). Integrating survey data with auxiliary sources of information to estimate crop yields. In JSM Proceedings, Survey Research Methods Section. Alexandria, VA: American Statistical Association.

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.

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.

Thank you!
Questions?

nathan.cruze@nass.usda.gov