A Bayesian Hierarchical Model for Combining Several Crop Yield Indications

Nathan B. Cruze

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

FCSM 2015
Washington, D.C.
December 1, 2015

Goal and technical approach

- **Goal:** Model sequence of in-season forecasts and estimates of crop yield
  - NASS Crop Production Report–state and national yield estimates
  - Reproducibility with appropriate measures of uncertainty
- **Approach:** Bayesian hierarchical model–synthesis of data from several surveys
  - Enforce physical relationships at two spatial scales
  - Incorporate variety of auxiliary data types

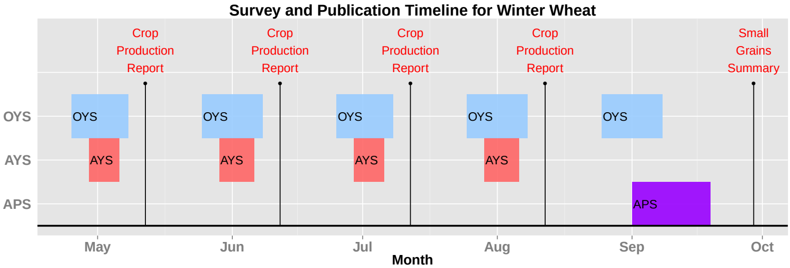
Challenge: From data to publication in 3-4 days

NASS crop yield surveys and reports

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

NASS surveys: Objective Yield (**OYS**), Agricultural Yield (**AYS**), Acreage, Production, and Stocks (**APS**)



Role of the Agricultural Statistics Board (ASB)

Expert panel of commodity specialists

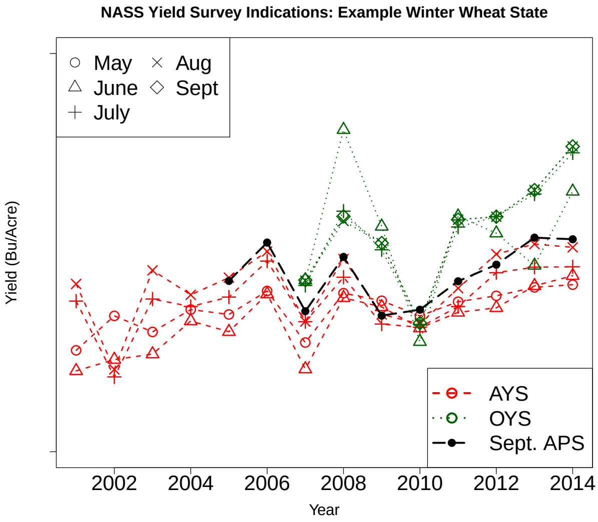
- Current and historical survey 'indications'
- Other information, e.g., weather, crop condition ratings
- Consensus on yield

Publish national and state estimates

OMB Standard 4.1 (2006): "Agencies must use accepted theory and methods when deriving...projections that use survey data. Error estimates must be calculated and disseminated to support assessment of the appropriateness of the uses of the estimates or projections..."

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

Example survey data



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

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

Region process model

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

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 for speculative region

Likelihood function–assuming conditional independence

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

Posterior distribution

$$[\mu_t, \boldsymbol{\Theta}_d, \boldsymbol{\Theta}_p | y_O, y_A, y_Q] \propto \prod_{k \in \{O,A,Q\}} [y_k | \mu_t, \boldsymbol{\Theta}_d][\mu | \boldsymbol{\Theta}_p][\boldsymbol{\Theta}_d][\boldsymbol{\Theta}_p] \tag{5}$$

Full conditional of regional yield, μ_t

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

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

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

Bayesian hierarchical model–state level yield

State-level counterparts indexed by $j \in \{1, 2, ..., J\}$

Unconstrained State Model–Define $\boldsymbol{\mu}_{t\cdot} \equiv (\mu_{t1}, \mu_{t2}, \ldots, \mu_{tJ})$,

$$\boldsymbol{\mu}_{t\cdot} | \boldsymbol{y}, \boldsymbol{\Theta}_d, \boldsymbol{\Theta}_p, \sim \textit{indep MVN} \left(\textit{vec} \left(\frac{\Delta_{2j}}{\Delta_{1j}} \right), \textit{diag} \left(\frac{1}{\Delta_{1j}} \right) \right) \tag{9}$$

Constrained State Model–Enforce constraint by conditioning (9) on $\mu_t = \sum_j w_j \mu_{tj}$

$$\left(\mu_{t1}, \mu_{t2}, \ldots, \mu_{t(J-1)} \right) \sim MVN(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\Sigma}}) \tag{10}$$

$$\mu_{tJ} = \mu_t - \frac{1}{w_{tJ}} \sum_{j=1}^{J-1} w_{tj} \mu_{tj} \tag{11}$$

Summary of model outputs

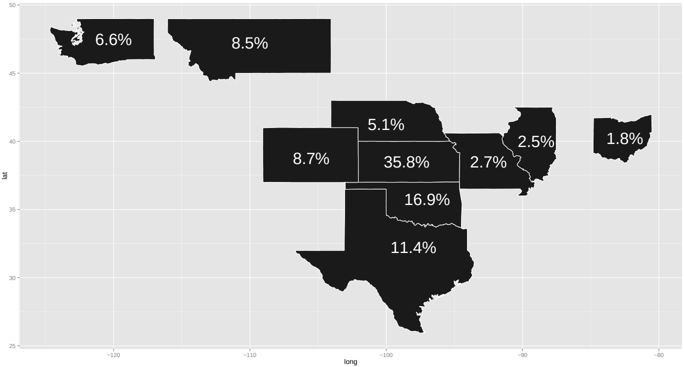
Speculative Region Model	Constrained State Model	Unconstrained State Model
Region yield and error Region forecast decomposition	Benchmarked state yields and errors	State forecast decompositions and benchmarking adjustments
Wang et al. (2012)	Adrian (2012), Nandram et al. (2014),Cruze (2015)	Kass and Steffey (1989)

		State 1	State 2	$\cdots$	State J	SPEC
Overall Forecast	$\hat{\mu}_{Tj}$	×	×	$\cdots$	×	×
Error		×	×	$\cdots$	×	×
OYS	$y_{OTm*j} - \hat{b}_{Om*}$	×	×	$\cdots$	×	×
AYS	$y_{ATm*j} - \hat{b}_{Am*}$	×	×	$\cdots$	×	×
Covariates	$\boldsymbol{z}_T'\hat{\boldsymbol{\beta}}$	×	×	$\cdots$	×	×
Sept. APS	y_{QTj}	×	×	$\cdots$	×	×
Benchmarking Adj.	d_j	×	×	$\cdots$	×	

$$\hat{\mu}_{tj} \approx \sum_{k \in \{O,A,Q,Covariates\}} c_k (SOURCE)_k + d_j \tag{12}$$

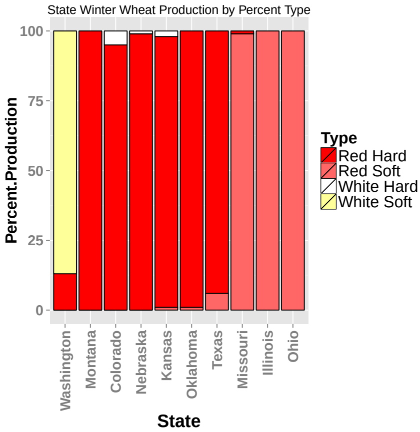
$$c_k \propto (variance)_k^{-1}$$

Winter wheat speculative region



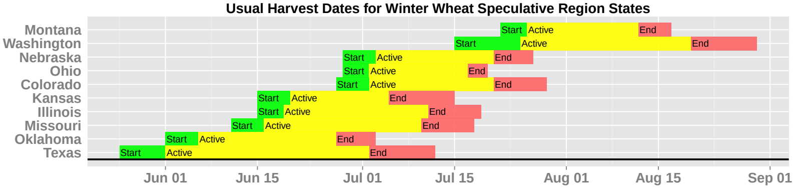
- 10 state region–some states geographically isolated
- Kansas has major share of harvested acres (Plotted: w_j, 2012)
- Four distinct types of winter wheat
- Differential planting and harvest

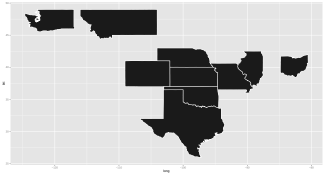
Winter wheat speculative region–types of wheat



- States 'specialize'
- Soft varieties associated with higher yield
- Washington, Missouri, Illinois, Ohio have higher yields
- Confounding with state

Winter wheat speculative region–differential harvest





- May OYS: only TX, OK, KS
- Southern states complete harvest before northern states begin
- Timing of covariates
- Deriving covariates for the region

Winter wheat model–covariates

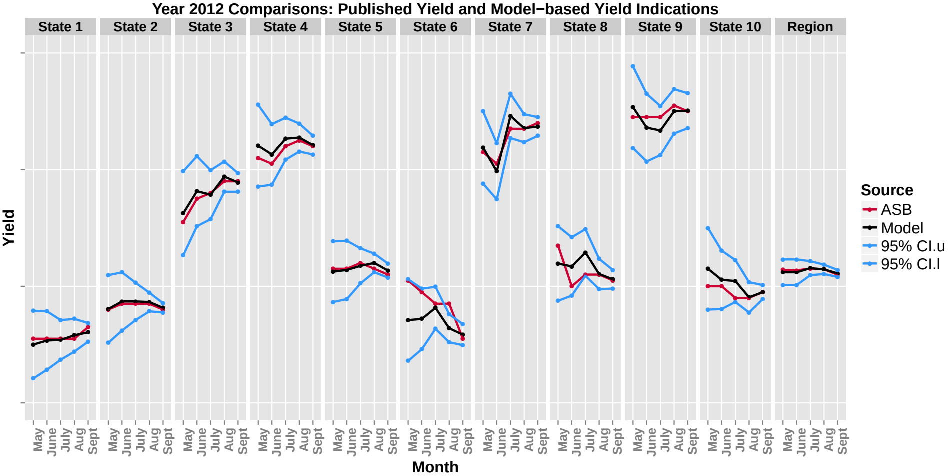
Covariates reflect conditions approaching active harvest dates

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

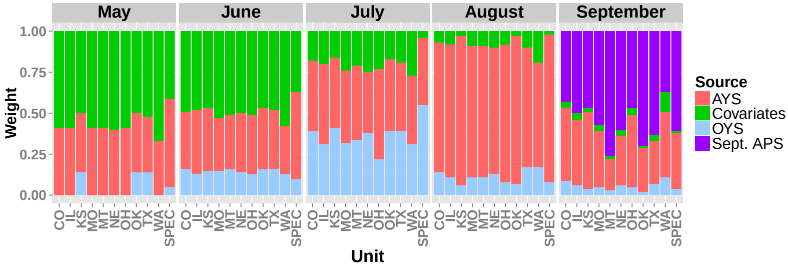
- State-specific constant
- z_{j2}: Linear time trend
- z_{j3}: Monthly precipitation (NOAA)
- z_{j4}: Monthly avg. temperature (NOAA)
- z_{j5}: Crop condition–% good + % excellent week # (NASS)

State/FIPS		**May Covars**		**June Covars**		**July-September Covars**	
		Condition (Week #)	*Weather (Month)*	*Condition (Week #)*	*Weather (Month)*	*Condition (Week #)*	*Weather (Month)*
CO	8	15	April	21	May	21	May
IL	17	15	April	19	May	19	May
KS	20	15	April	19	May	19	May
MO	29	15	April	19	May	19	May
MT	30	15	April	19	May	24	June
NE	31	15	April	21	May	21	May
OH	39	15	April	21	May	21	May
OK	40	15	April	17	April	17	April
TX	48	15	April	17	April	17	April
WA	53	15	April	22	May	22	May

Comparing ASB estimates and model outputs–2012



Weights applied in wheat forecast decomposition



- Early season emphasis on covariates
- Increasing emphasis on OYS in July
- Heavy emphasis on last AYS in August
- Heavy emphasis on quarterly survey in September

Extensions and conclusions

1. NASS yield models (corn, soybeans, winter wheat) capture expert assessment in manner which is reproducible and provide justifiable measures of uncertainty.
2. This methodology is flexible enough to accommodate many types of auxiliary data.

- Additional commodities
- Non-spec region states
- New technologies, e.g., soil moisture monitors

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.

Kass, R. and Steffey, D. (1989). Approximate Bayesian inference in conditionally independent hierarchical models (parametric empirical Bayes models). *Journal of the American Statistical Association*, 84(407):717–726.

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