A Bayesian Hierarchical Model for Combining Several Crop Yield Indications

Nathan B. Cruze*

Abstract

USDA's National Agricultural Statistics Service (NASS) conducts multiple surveys over the course of a growing season. Each of these surveys reflects current growing conditions and provides a prediction of end-of-season crop yield. In particular, NASS conducts two interview-based surveys and one field measurement survey from which indications of crop yield may be obtained. It is also known that a number of weather conditions during the growing season may contribute to changes in crop yield. This paper describes a Bayesian hierarchical model that improves end-of-season crop yield predictions by combining these several disparate sources of information. The model incorporates benchmarking of state-level forecasts with regional forecasts of crop yield and gives rise to rigorous measures of uncertainty. It also permits a useful decomposition with respect to the emphasis placed on each information source.

Key Words: Bayesian hierarchical model; Composite estimation; Model-based estimation; Survey sampling

1. Introduction

The mission of USDA's National Agricultural Statistics Service (NASS) is to provide timely, accurate, and useful statistics in service to U.S. agriculture. NASS publishes within-season forecasts of state and national harvested acreage totals, production totals, and yield per area harvested in its monthly Crop Production Report. These official statistics reflect the consensus estimates agreed upon by NASS's Agricultural Statistics Board (ASB) after review of current and historical survey outcomes, administrative data, and other relevant information on weather and crop condition.

Due to the importance of these official agricultural statistics in informing commodity market expectations, NASS has an ongoing interest in strengthening its traditional estimation procedures. This paper outlines a model-based procedure for estimating state and regional crop yields. The input data sources and requirements of the NASS yield forecasting program are outlined in section 2. The proposed methodology in section 3 details a Bayesian hierarchical model that combines several distinct survey inputs, as well as auxiliary information to produce benchmarked, one-number forecasts of crop yield at state and regional levels. The model offers an easily reproducible means of estimating crop yield given possibly disparate sources of information while providing rigorous measures of uncertainty. Some empirical results for winter wheat are discussed in section 4; the yield models for winter wheat are shown to perform well over a wide variety of conditions. Discussion and conclusions are offered in section 5.

2. NASS Crop Yield Surveys and the Monthly Crop Production Report

The creation and dissemination of the NASS Crop Production Report has a long history dating back to statutes codified in 1909. Specifically, 7 USC Sec. 411a, describes the necessity of monthly crop reports, as well as the contents, issuance and approval by the

*USDA National Agricultural Statistics Service (NASS), Room 6412A–South Building, 1400 Independence Ave., SW, Washington, DC 20250

Secretary of Agriculture. Furthermore, this code dictates that the Crop Production Report "...shall be printed and distributed on or before the twelfth day of each month." (Allen, 2007, p. 19) Presently, NASS supports official in-season forecasts and estimates of state and national crop yield for its major small grains and row crops with a 5-month survey cycle comprised of three probability-based surveys. The survey cycle and approximate data collection windows for winter wheat are depicted in Figure 1 where the three surveys are the Objective Yield Survey (OYS), Agricultural Yield Survey (AYS), and the quarterly Acreage, Production and Stocks (APS) Survey.

- The OYS is a monthly survey based on *field measurements* obtained at plots of land throughout the season. The survey is commodity specific (currently conducted only for corn, soybeans, winter wheat, cotton and potatoes). Due to its considerable expense, it is conducted only in a limited region known as the *speculative region*, a group of major producing states as determined by NASS.

- The AYS is a monthly interview-based survey. The AYS survey is designed to provide coverage for all small grains or row crops within the growing season, and it is conducted nationwide.

- Like the AYS, the quarterly APS survey is an interview-based survey. It is conducted with a much larger sample size than either the AYS or the OYS, and it is used to obtain indications of changes in stocks, planted and harvested area, and production in addition to yield. Since it is conducted *post-harvest* when the weather events and decisions of the current crop year have been fully realized, the APS yield indication is generally considered the 'gold standard' of all NASS surveys.

Each survey is conducted with a first-of-month reference date. Row crops including corn and soybeans are supported by a similar timeline between August and December. The NASS production timeline generally permits a three or four day window between the summary of all surveys and the release of the official yield forecast in the Crop Production Report no later than the twelfth day of the month. Final wheat yield estimates are released in the Small Grains Summary in late September.

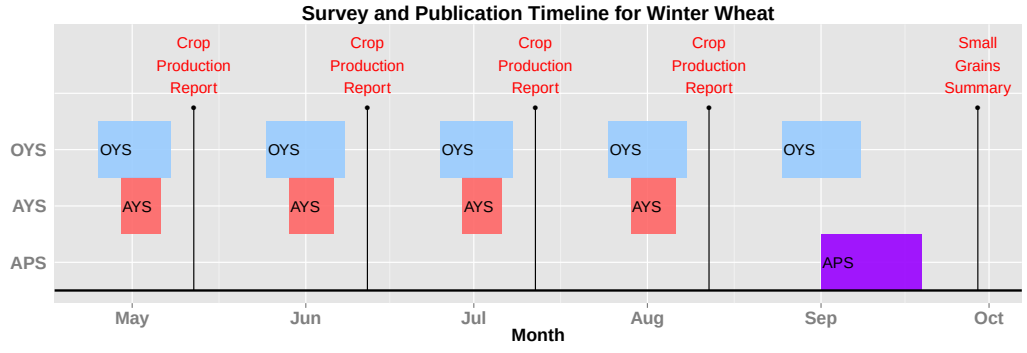

Figure 1: Survey and report production timeline for NASS winter wheat forecasts and estimates

In any given month, more than one survey *indication* of crop yield is available. Survey indications and standard errors are available for states and for the speculative region for all three surveys. (The AYS and APS also supply national indications.) During deliberations, members of the ASB review current and historical survey indications, and consider other information on crop condition and weather. In the interest of not disclosing NASS

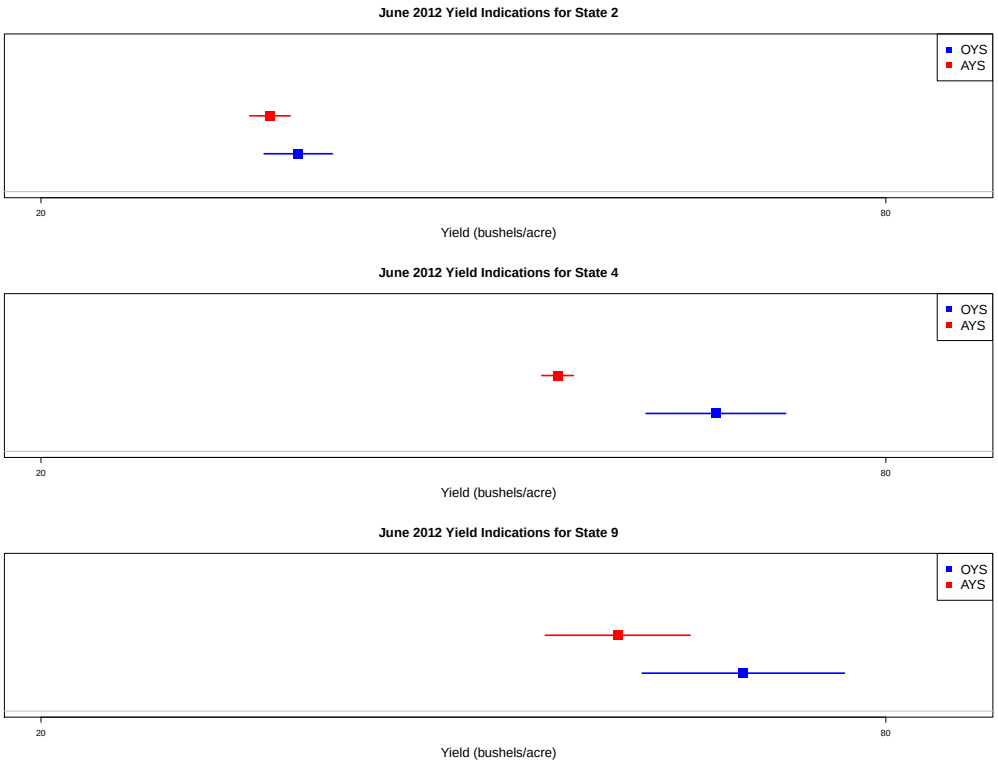

Figure 2: June 2012 Objective Yield and Agricultural Yield Survey indications and 95% confidence intervals for select states in the winter wheat speculative region

indications, state names have been redacted and replaced with an arbitrarily assigned index number to be used consistently through out the remainder of this paper. Figure 2 shows OYS and AYS indications and 95% confidence intervals for three states in the month of June during the 2012 crop year. The figure shows that yield can vary dramatically by state, that the corresponding yield indications may be disparate, and that the standard errors can differ by state and by survey. ASB experience shows that the OYS indication is generally *biased upward* relative to the end-of-season APS yield indication. By contrast, the AYS indications are *biased downward*. The extent of these biases may decrease with each month as conditions of the growing season are more fully realized. The relationships between AYS, OYS, and APS indications generally hold irrespective of commodity.

One important aspect of the ASB deliberation is setting a one-number yield estimate for the speculative region. While NASS does not publish this in its Crop Production Report, the ASB members must reach a consensus estimate for the speculative region yield given the available survey and auxiliary data. Pursuant to the speculative region estimate, estimates for member states are set given the survey and auxiliary data, subject to the constraint that regional yield, denoted μ, is equal to a weighted average of the yields of member states μ_j, where the weight w_j is determined in proportion to the j^{th} state's harvested area.

$$\mu = \sum_{j=1}^{J} w_j \mu_j \tag{1}$$

The identity in Equation 1 follows from observing that total output and total harvested area for the region are sums of those totals at the state level, and in some sense, it embodies a balance of materials, both in terms of output and total area harvested. Therefore, any official statistics for yield should satisfy Equation 1.

3. Bayesian Hierarchical Model for Combining Survey Indications

The NASS official statistics are the result of the expert assessment of the ASB. While the ASB can evaluate many sources of information and react accordingly, it is not simple to disclose the reasons for their adjustments, or which information source has been trusted in setting the official statistics. Moreover, the ASB process does not give rise to measures of uncertainty. The proposed value-added of modeling the ASB process is that it might make the ASB process more easily reproducible, provide some interpretation of the role of various input information sources, and produce estimates of uncertainty.

The Bayesian hierarchical model outlined below has evolved from the initial research of Wang et al. (2012) in which the problem of combining survey and auxiliary information was considered exclusively at the speculative region level for corn and soybeans. Subsequent work by Nandram et al. (2014) introduced benchmarking of estimates of member states to the speculative region yield. The work of Adrian (2012) introduced further simplification to the models at both regional and state scale, and it informed the work of Cruze (2015) and the present work on winter wheat yield.

The ASB has received model-based indications of corn and soybean yields for their deliberation since 2011. An operational winter wheat yield model has been provided to the ASB since 2015. In this section, a general methodology for NASS current practice is presented. Empirical results for winter wheat are provided in section 4.

3.1 Models for the Speculative Region

As in Wikle (2003) and others, the Bayesian hierarchical model can be specified as a collection of conditional and marginal distributions in three parts: a data model that describes the behavior of observed data given some underlying process for yield, the process model that relates the latent yield (the parameter of interest, denoted μ_t) to observable covariates, and prior distributions for model parameters. Let y_{ktm} denote observed yield indications from survey $k \in \{O, A, Q\}$ (for OYS, AYS, and quarterly APS, respectively), in year $t \in \{1, 2, ..., T\}$ and month m. Conditional on the latent regional yield, μ_t, data models for forecast month $m*$ are described by

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

and

$$y_{Qt}|\mu_t \sim \ indep\ N\left(\mu_t, s^2_{Qt}\right) \tag{3}$$

where it is understood in Equation 3 that the survey is conducted in September for winter wheat. Equation 2 states that, conditional on the yield process, the AYS and OYS monthly indications are observations of true yield with month-specific biases b_{km*}. The APS survey indication is assumed to be an unbiased indication for true yield.

The process model describes variation of true end-of-season yield μ_t about a linear function of covariates, $\boldsymbol{z}_t$.

$$\mu_t \sim \ indep\ N\left(\boldsymbol{z}_t'\boldsymbol{\beta}, \sigma^2_\eta\right) \tag{4}$$

Finally, diffuse prior distributions complete the specification of model; for b_{km*} and $\boldsymbol{\beta} \sim indep\ N(0, 10^6)$ and σ^2_{km},$\sigma^2_\eta \sim indep\ IG(.001, .001)$. As a convenience, the collection of data model parameters will be denoted $\boldsymbol{\Theta}_d \equiv \left(b_{km*}, \sigma^2_{km*}\right)$ and the vector of process model parameters $\boldsymbol{\Theta}_p \equiv \left(\boldsymbol{\beta}, \sigma^2_\eta\right)$.

Assuming conditional independence, the likelihood function has the following form

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

and it follows by Bayes' Rule that the posterior distribution takes the form shown in Equation 6:

$$[\mu_t, \boldsymbol{\Theta}_d, \boldsymbol{\Theta}_p | y_O, y_A, y_Q] \propto \prod_{k \in \{O,A,Q\}} [y_k | \mu_t, \boldsymbol{\Theta}_d][\mu | \boldsymbol{\Theta}_p][\boldsymbol{\Theta}_d][\boldsymbol{\Theta}_p] \tag{6}$$

A Gibbs sampling algorithm Gelman et al. (2003) is employed to obtain estimates of all model parameters. For brevity, we discuss only the full conditional distribution for regional yield μ_t,

$$[\mu_t | y_O, y_A, y_Q, \boldsymbol{\Theta}_d, \boldsymbol{\Theta}_p] \sim N\left(\frac{\Delta_2}{\Delta_1}, \frac{1}{\Delta_1}\right) \tag{7}$$

where

$$\Delta_1 = \sum_{k=O,A} \frac{1}{\sigma^2_{km*} + s^2_{ktm*}} + \frac{I_{\{Q\}}}{s^2_{Qt}} + \frac{1}{\sigma^2_\eta} \tag{8}$$

$$\Delta_2 = \sum_{k=O,A} \frac{y_{ktm*} - b_{km*}}{\sigma^2_{km*} + s^2_{ktm*}} + \frac{I_{\{Q\}} y_{Qt}}{s^2_{Qt}} + \frac{\boldsymbol{z}'_t \boldsymbol{\beta}}{\sigma^2_\eta}. \tag{9}$$

Equation 8 describes the sum of the precisions of each information source. Dividing Equation 9 by Equation 8, the mean of the full conditional distribution Equation 7 is *shown to be a weighted average of available information sources*: the bias-corrected AYS and OYS indications, the quarterly APS indication (when it is available), and covariates information. This relationship serves as a useful interpretation for the one number yield forecast as a *meaningful composite* of the available information, and the most precise information sources receive a proportionally larger weight in the overall yield.

3.2 Models for States

Data and process models for the states resemble those of the speculative region with models for each state j given by:

$$y_{ktm^*j} | \mu_{tj} \sim indep\ N\left(\mu_{tj} + b_{km^*j}, s^2_{ktm^*j} + \sigma^2_{km^*j}\right), k = O, A, \tag{10}$$

$$y_{Qtj} | \mu_{tj} \sim indep\ N\left(\mu_{tj}, s^2_{Qtj}\right), \tag{11}$$

$$\mu_{tj} \sim indep\ N\left(\boldsymbol{z}'_{tj}\boldsymbol{\beta}_j, \sigma^2_{\eta j}\right). \tag{12}$$

Diffuse prior distributions are specified on the data and process model parameters of each state as before. The full conditional distribution of yield in the j^{th} state, μ_{tj} resembles Equation 7. Assuming independence, the collection of state-level crop yields follows a multivariate normal distribution.

$$\boldsymbol{\mu}_{t\cdot} | \boldsymbol{y}, \boldsymbol{\Theta}_d, \boldsymbol{\Theta}_p, \sim indep\ MVN\left(vec\left(\frac{\Delta_{2tj}}{\Delta_{1tj}}\right), diag\left(\frac{1}{\Delta_{1tj}}\right)\right) \tag{13}$$

While parameters μ_{tj} must respect the balance identity in Equation 1, estimates of parameters $\hat{\mu}_{tj}$ derived under Equation 13 may not. Therefore, it is desirable to enforce the balance constraint between the speculative region and member states. Iterates of the speculative region MCMC simulation are fed into the MCMC simulation for a 'constrained' state level model. By conditioning the vector of state-level yields in Equation 13 on the speculative region yield μ_t, the collection of the first $j-1$ states will follow a multivariate normal distribution

$$\left(\mu_{t1}, \mu_{t2}, \ldots, \mu_{t(J-1)}\right) \sim MVN(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\Sigma}}). \tag{14}$$

At each time t, the yield for the J^{th} state is given by

$$\mu_{tJ} = \mu_t - \frac{1}{w_{tJ}} \sum_{j=1}^{J-1} w_{tj}\mu_{tj}, \tag{15}$$

which resembles the top-down procedure used during the ASB's own decision making process. Posterior means obtained from the Monte Carlo samples under Equation 7, Equation 14, and Equation 15 a collection of point estimates for the speculative region and all member states that honor Equation 1. Posterior variances serve as standard errors with these estimates, giving rise to defensible measures of uncertainty at both spatial scales.

4. Model-based Estimates of Winter Wheat Yield

The winter wheat speculative region consists of ten states: Colorado, Illinois, Kansas, Missouri, Montana, Nebraska, Ohio, Oklahoma, Texas, and Washington. NASS identifies these states as producers of winter wheat representative of one of four classes of winter wheat: *hard red, hard white, soft red, or soft white winter wheat*. Generally, soft varieties of winter wheat tend to show higher yields, and states that specialize in those classes also tend to show remarkably higher yields.

Because of the size and geographic spread of this region, the harvest is initiated at different times within the yield forecasting season. Differential harvest tends to start early in the south and later in northern states as depicted in Figure 3. The differential growth and development of the crop affects the availability of some indications. In particular, only Texas, Oklahoma, and Kansas participate in the May Objective Yield Survey. The remaining seven states join the OYS sample from June onward. All 10 states participate in the AYS from May through August.

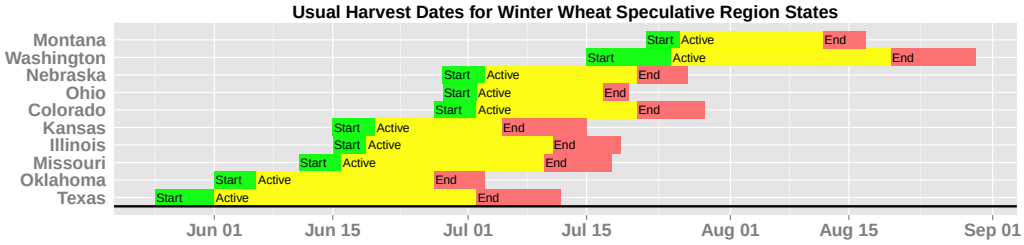

Figure 3: Typical harvest ranges for winter wheat speculative region states

Differential harvest also informs the selection of covariates. The fitted process model for each state has been taken as

$$\hat{\mu}_{tj} = \hat{\beta}_{j1} + \hat{\beta}_{j2}TREND_{j2} + \hat{\beta}_{j3}PCP_{j3} + \hat{\beta}_{j4}TEMP_{j4} + \hat{\beta}_{j5}CONDITION_{j5} \tag{16}$$

where $TREND$ is a linear trend term to capture any innovation in yield over time, PCP is the state's monthly precipitation, $TEMP$ is the monthly average temperature and the variable $CONDITION$ is the percent of the crop that has been rated good or excellent according to NASS's crop condition ratings. The May model for each state is initialized with weather outcomes and crop condition ratings observable near the May 1 reference date. Weather and crop condition variables are updated for select states reflecting the onset of harvest shown in Figure 3. The same covariates at the speculative region level are derived

from a weighted average of state covariates, weighted in proportion to each state's share of harvested area.

The specified model has been used by the ASB only as of the 2015 crop year. Figure 4 depicts the sequence of model-based yield forecasts and final estimates during the 2012 crop year. For comparison, NASS official statistics (the expert assessment of the ASB) are shown in red. In a year in which the model was unavailable to inform ASB opinion in any way, the model seems to capture the expert assessment of the ASB very well. The official statistics are generally well within the 95% credible intervals of the model-based estimate, and the model trues up well by season's end.

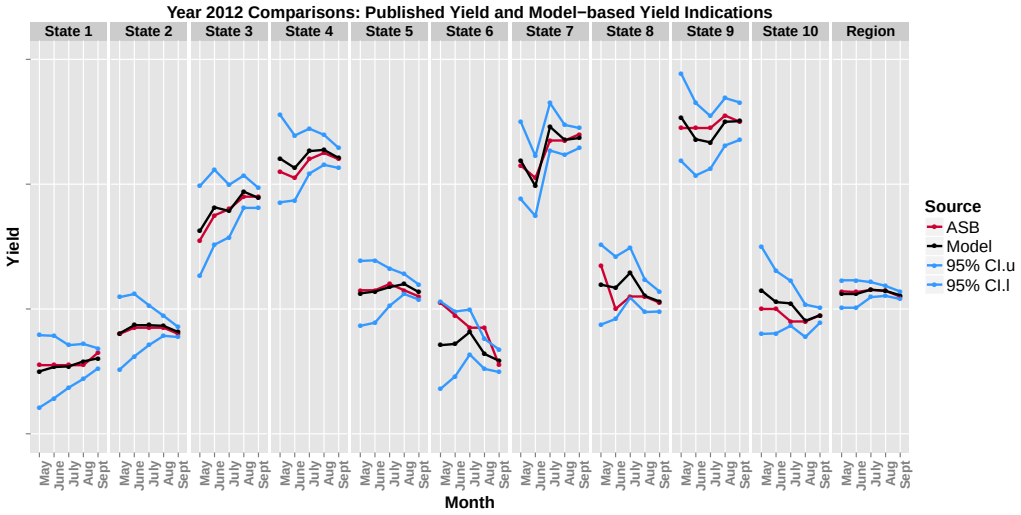

Figure 4: Performance of regional and state model-based forecasts and estimates compared to NASS official statistics for winter wheat yields for 10 speculative region states.

A salient feature of the model is the decomposition of the overall state and regional yield forecasts by information source. Both state and regional forecasts may be interpreted approximately as weighted averages of the input information sources. For the 2012 crop year, the emphasis applied to each information source in the model-based forecasts is shown in Figure 5 to vary by month. Early in the growing season, the regression component incorporating chosen covariates receives the heaviest emphasis. As the events of season are realized, the emphasis shifts from covariates, to bias-adjusted OYS indications (July), bias-corrected AYS indications (August) and the gold-standard APS survey indications (September).

The posterior distributions under the constrained state model also afford a way to assess 'how likely' each separate state indication is in terms of a posterior probability. This is illustrated graphically for the select states shown in Figure 6 in reference to June of the 2012 crop year. The raw OYS and AYS survey indications (shown as squares) are closer to the tails of the posterior distributions. Ideally, the bias correction should not just address the upward (downward) tendency of the OYS (AYS) indications, but it should also *reduce the spread* between the input information sources. These input information sources fall into regions of highest posterior density, and the posterior mean, shown in black, may be interpreted approximately as a composite of those information sources. From Figure 5, the emphasis was shown to lie heavily on the covariates in June. This is born out by the proximity of the posterior means relative to the covariates component of the model.

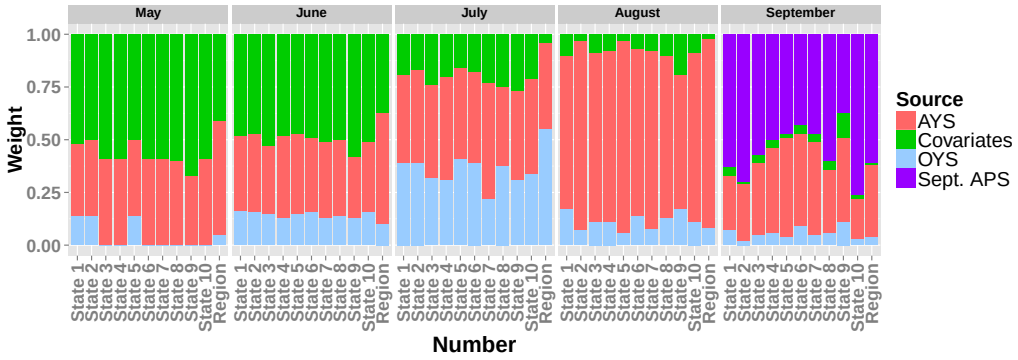

Figure 5: Approximate weight applied to each information source under the proposed model

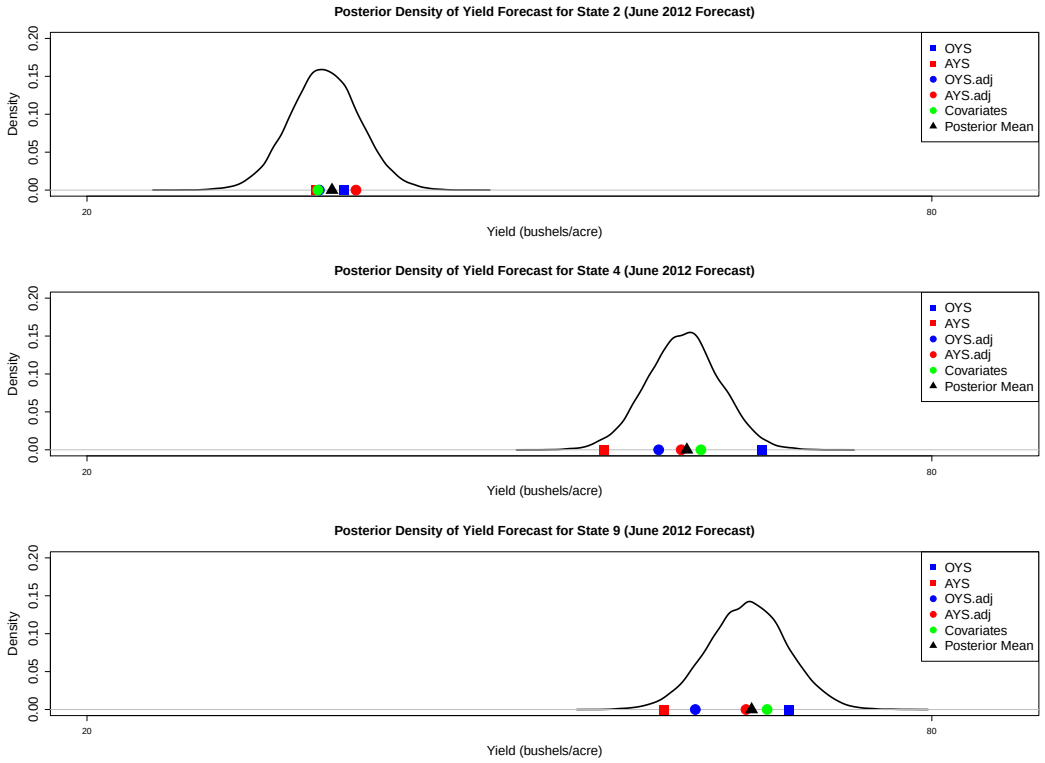

Figure 6: Comparison of indications under the constrained state model for select states

5. Discussion and Conclusions

The proposed methodology has been used by NASS in recent years to provide another useful indication to ASB decision makers in support of estimates of corn, soybeans, and winter wheat. An ongoing research challenge is to produce models that perform well over the wide variety of year-to-year planting decisions and anomalous weather conditions and events. Figure 7 shows the year-over-year performance of the model in a southern state that experienced a 50% decline in yield between 2012 and 2014. The specified model for winter wheat captures that decline, and matches the expertise of the ASB very closely. Continued refinements to the crop yield models may incorporate additional covariates as

new information and new technologies become available. Making use of the same types of information at finer temporal or spatial scales may also help robustify the model against anomalous conditions.

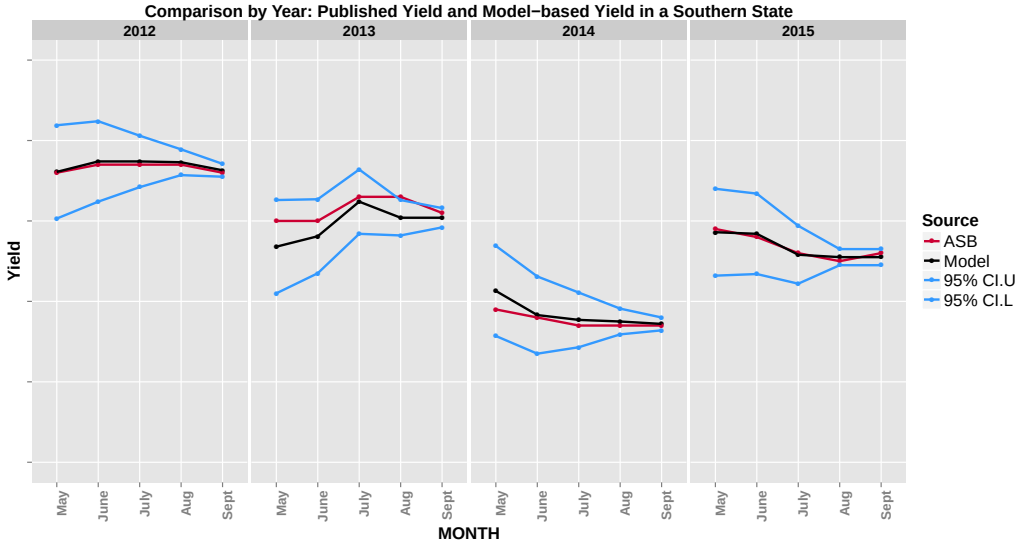

Figure 7: The specified model captures year-to-year differences in yield in this southern state

The proposed methodology is flexible, and it can be adapted for the use for other commodities. Corn, soybeans, and winter wheat represent only a few of the commodities covered in the NASS Crop Production Report. Future work will look to extend the existing model to other commodities such as upland cotton. Adapting the model for coverage for non-speculative region states (which do not participate in the Objective Yield Survey) and the national program is another important goal as NASS seeks to strengthen its crop yield forecasting program with model-based strategies.

References

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

Allen, R. (2007). Safeguarding america's agricultural statistics: A century of successful and secure procedures, 1905-2005. USDA National Agricultural Statistics Service.

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.

Gelman, A., Carlin, J., Stern, H., and Rubin, D. (2003). *Bayesian Data Analysis (2nd ed.)*. Chapman & Hall/CRC.

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.

Nandram, B. and Sayit, H. (2011). A Bayesian analysis of small area probabilities under a constraint. *Survey Methodology*, 37:137–152.

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.

Wikle, C. (2003). Hierarchical Bayesian Models for Predicting the Spread of Ecological Processes. *Ecology*, 84:1382–1394.