A Model-Based Approach to Crop Yield Forecasting

Nathan B. Cruze[*]
Habtamu K. Benecha

Abstract
The USDA's National Agricultural Statistics Service conducts multiple surveys for major crops during the growing season. These surveys are designed to capture the current status of crops at state, regional, and national levels with a first-of-the-month reference date. Each of the surveys also provides an estimate of potential end-of-season crop yield. We extend a Bayesian hierarchical model to produce improved yield forecasts for upland cotton. The model combines these possibly disparate survey estimates together with auxiliary data to produce one-number forecasts for a region and its member states. The resulting state forecasts are benchmarked against the regional forecast. The model gives rise to easily reproducible estimates with rigorous measures of uncertainty. The proposed candidate model for upland cotton is shown to perform well over a wide variety of growing conditions. Some particular challenges of modeling upland cotton are noted.

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. As a federal statistical agency, NASS is compelled by law to publish a monthly Crop Production Report no later than the twelfth day of each month. While the contents of the Crop Production Report will vary reflecting the seasonality of agriculture in the United States, some of the key contents are its within-season forecasts of three related quantities at both the national and state levels: harvested acreage totals, projected total production, and forecasted crop yield (total production per area harvested). 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. Official statistics generated in this manner do not give rise to measures of uncertainty, therefore, limited uncertainty information is provided with the published statistics in the Crop Production Report.

Since 2011, NASS has used model-based forecasts in tandem with its traditional estimation procedures. Modeled forecats for commodity specific regions and member states have been provided as additional inputs into the ASB's deliberations for corn, soybeans, and winter wheat. At the request of its ASB stakeholders, NASS began researching forecasting techniques for upland cotton yield. This paper outlines a model-based procedure for estimating state and regional crop yields for upland cotton. The input surveys and requirements of the NASS yield forecasting program are discussed in Section 2. The proposed methodology in Section 3 details a Bayesian hierarchical model that combines several distinct survey inputs and available auxiliary information to produce benchmarked, one-number forecasts of crop yield at state and regional levels. In the long run, models of this type may offer an easily reproducible means of estimating crop yield given possibly disparate sources of information while providing rigorous measures of uncertainty. The results of one candidate

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

model for upland cotton are presented in Section 4. Discussion and preliminary conclusions are provided in Section 5.

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

2.1 Necessity and Timing

Federal law (7 USC Sec. 411a) describes the necessity of monthly crop reports, and their contents, issuance and ultimate approval by the Secretary of Agriculture (Allen, 2007, p. 19). NASS crop yield forecasts provide an important source of information to commodities markets and help inform farm policy. Presently, NASS supports these important official in-season forecasts of state and national crop yield for its major row crops over a six month cycle beginning in August and continuing through the release of its Crops Annual Summary in early January of the following calendar year. In order to produce this report, NASS conducts three probability-based surveys: the Objective Yield Survey (OYS), the Agricultural Yield Survey (AYS), and the December Quarterly Acreage, Production and Stocks (APS) Survey. For upland cotton in particular, these surveys provide coverage for the 17 southern states shown in Figure 1 and a six state subregion called the *speculative region*.

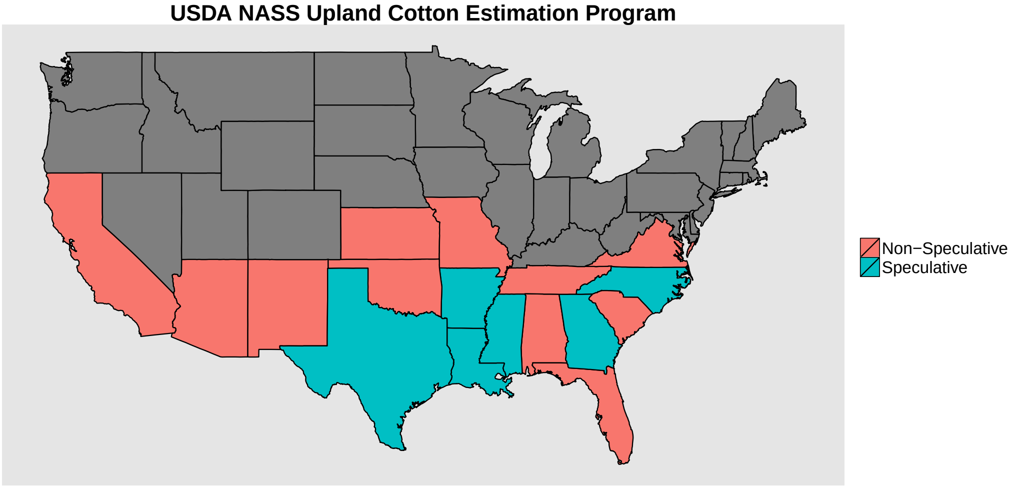


Figure 1: USDA NASS Upland Cotton Estimation Program States and Speculative Region

The OYS is a monthly survey based on *field measurements* obtained at sampled fields throughout the growing season. The survey is commodity specific. Currently, these surveys are only conducted for corn, soybeans, winter wheat, potatoes, and upland cotton. Due to its considerable expense, it is conducted *only* in the speculative region, a commodity-specific region comprised of some of the top producing states. The upland cotton speculative region is composed of Arkansas, Georgia, Louisiana, Mississippi, North Carolina, and Texas. From this survey, estimates and standard errors of cotton yield can be obtained at the regional level and for each of the six member states.

The AYS is a monthly interview-based survey. The AYS survey is designed to provide coverage for all major row crops within the growing season, and it is conducted nationwide each month from August through November. Among other questions, respondents are asked to provide their best assessment of final yield as of the first-of-month reference date.

Estimates and standard errors can be obtained for participating states, speculative regions, and the national program.

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 estimates of changes in stocks, planted and harvested area, and production in addition to yield. The survey produces a multiple frame estimate, with list frame undercoverage correction obtained from respondents identified in NASS's June Area Survey. Since the APS is conducted *post-harvest* when the weather events and decisions of the current crop year have been fully realized, it provides sound estimates and standard errors of end-of-season yield at state, regional, national levels.

As a result of the marketing processes surrounding upland cotton, a fourth source of information comes from a bi-weekly *census* of cotton gins (processors). The cotton ginnings (CG) projections provide a measure of projected total production for each of the 17 major cotton producing states, and it can be aggregated to regional or national program level. Cotton ginnings continue to evolve beyond the publication of the NASS Crops Annual Summary report. Ultimately, the ginnings totals are finalized in May of the following calendar year; in principle, every bale of cotton grown in the United States will be accounted for at this time. Coupled with a final estimate of harvested area, final cotton crop yield is known with near certainty in May, therefore NASS views May ginnings as the final estimate and the gold standard. Essentially, this is the quantity to be forecasted for each state, region, and the nation.

The NASS survey and publication timeline is shown in Figure 2, with the width of each box representing the approximate data collection window for each survey. For simplicity, first-of-month cotton ginnings are shown from October to January; October is generally the first month in which all states will begin reporting. The timeline illustrates that in any given month, at least two (possibly disparate) estimates can be obtained for the same quantity. The ASB convenes to synthesize the results of these probability surveys and cotton ginnings projections into its official one-number forecasts released in the monthly Crop Production Report, generally just three to four days after data collection has concluded. A sequence of six forecasts for upland cotton is made throughout the growing season, beginning in August and culminating in the release of preliminary annual estimates in January in the Crops Annual Summary.

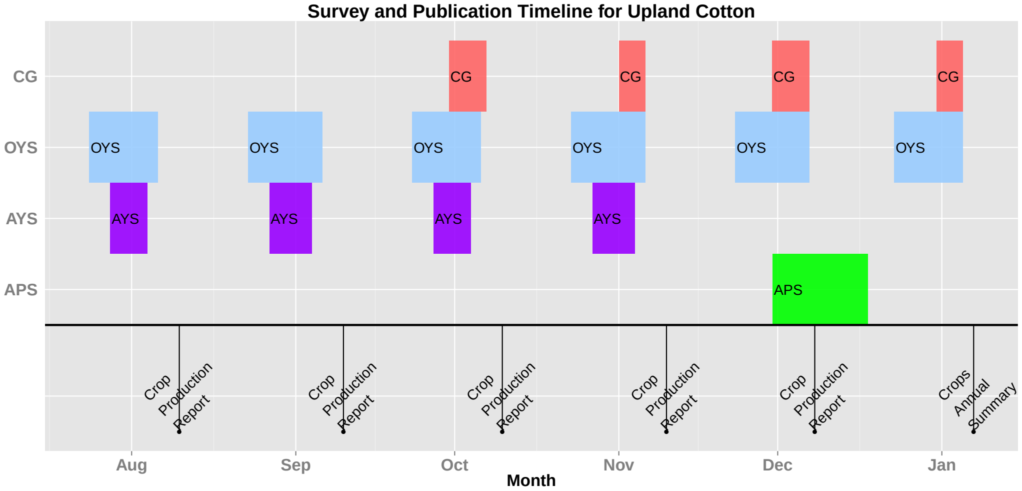
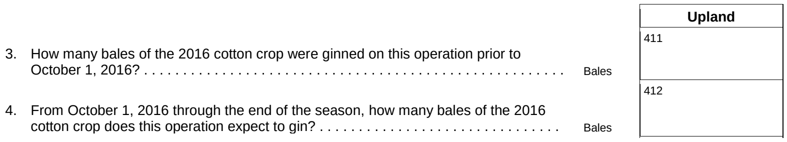


Figure 2: Survey and report production timeline for NASS upland cotton yield forecasts

2.2 Availability and Features of Input Estimates

2.2.1 Cotton Ginnings

Incorporating the projected cotton ginnings into a model for yield is desirable, but two challenges exist. First, the NASS cotton ginnings reports collect information about projected total production as opposed to yield per acre. Unlike the NASS probability surveys which are subject to sampling errors, the ginnings reports represent a census of all cotton processors in the United States. The within season ginnings are still subject to error and variability, however. In particular, all cotton gins in the nation are asked to respond to the questions shown below in Figure 3 regarding activity that has already taken place at the cotton gin as of the reference date, and an anticipated (forecasted) amount of activity that remains to take place. Despite the fact that the entire population of cotton gins is to be enumerated, the state totals are subject to forecasting errors. Additionally, nonresponse is possible; NASS accounts for this through the use of reweighted estimators. Thus, nonresponse results in an additional source of variability. Finally, corrective factors are applied to these totals to account for processing that may have taken place across state lines.



		Upland
3. How many bales of the 2016 cotton crop were ginned on this operation prior to October 1, 2016?	Bales	411
4. From October 1, 2016 through the end of the season, how many bales of the 2016 cotton crop does this operation expect to gin?	Bales	412

Figure 3: Questions regarding totals ginned and to be ginned

In order to account for these potential sources of variability and error, we propose an estimated mean square error. For each month, October through January, the estimated ginnings totals were converted to yield per acre by dividing total production by official NASS harvested area totals. These yields were compared to May ginnings per acre in the following calendar year, and the deviations were computed as illustrated for the the example state in Figure 4. In practice, May ginnings per acre cannot be known within the current year forecasting window. Thus, an average of squared deviations was computed using the previous 10 year history, e.g., the estimated October 2010 mean squared error was the average of October squared deviations from years 2000 to 2009. One desirable property of this estimator is that it reflects decreasing uncertainty as the season progresses, i.e., that October uncertainty is generally greater than January uncertainty. It also results in a history of ginnings per acre estimates and estimated means squared errors from 2008 to present, commensurate with the length of survey history available for the six state speculative region.

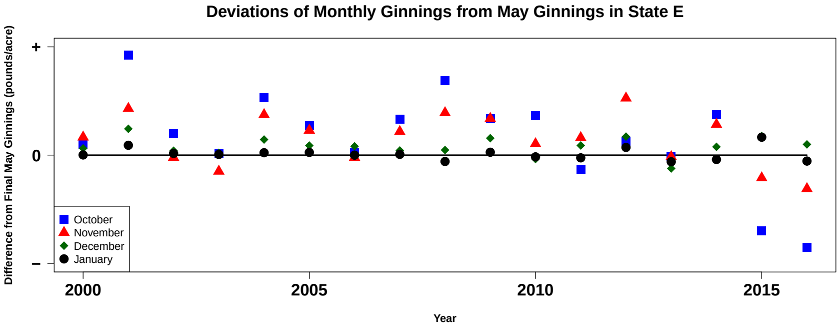


Figure 4: Deviations of ginnings per acre for example state

2.2.2 *Probability Survey Estimates versus the May Final Yields*

The six states included in the cotton speculative region have remained constant since 2008. (Prior to 2008 a seventh state was part of the region.) Figure 5 represents a complete history of OYS, AYS, and APS survey estimates for an example state in the speculative region. In any given month, more than one probability survey estimate of crop yield is available. Plotted relative to May ginnings, a bias tendency in each of these surveys is shown: OYS tends to be biased upward, AYS tend to be biased downward, and even the APS may show some small amount of bias relative to the full enumeration of all cotton produced by May. These tendencies generally hold in each state and even for different commodities. Thus, we think of the available survey estimates as potentially biased, with biases that may depend on forecast month.

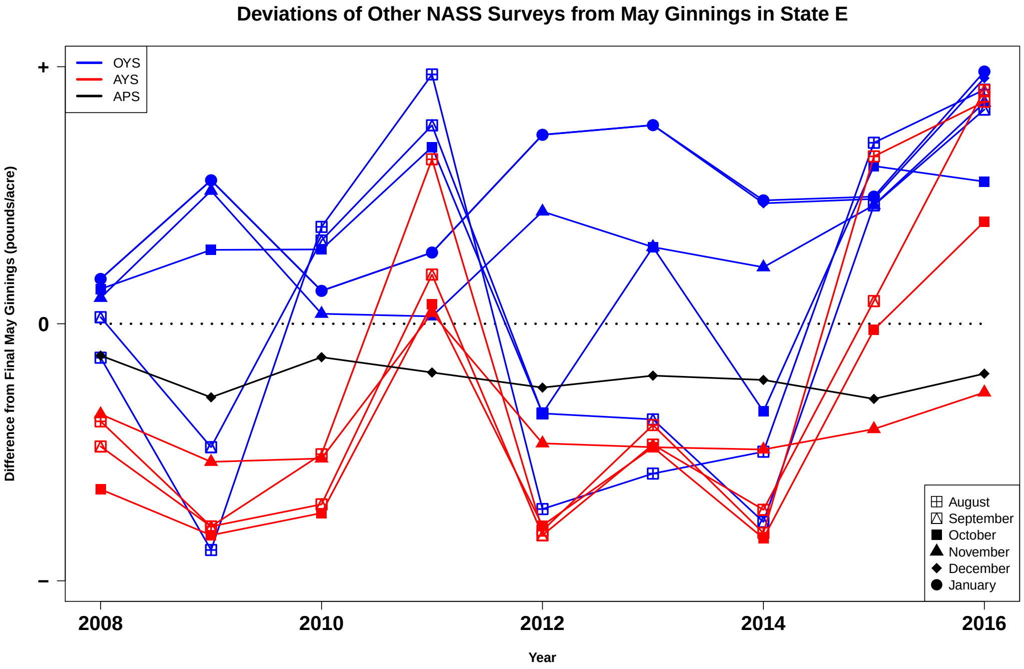


Figure 5: History of survey estimates relative to final May yields

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 how various sources of information have been utilized in setting the official statistics. Moreover, the ASB process does not give rise to measures of uncertainty. One benefit of modeling cotton yield 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 its roots in early research by 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 additional work on winter wheat Cruze (2015, 2016).

The ASB has received model-based indications of corn and soybean yields for their deliberation since 2011. An operational winter wheat yield model was provided to the ASB beginning in 2015. In this section, an extension to NASS's general methodology is presented which incorporates the additional cotton ginnings into model-based forecats of upland cotton. Empirical results for a candidate model for upland cotton are provided in Section 4.

3.1 Models for the Speculative Region

The literature on Bayesian hierarchical models and their application is vast. A common strategy, e.g., as in Wikle (2003), is to specify the Bayesian hierarchical model as a collection of conditional and marginal distributions in three parts: a data model that describes the behavior of observed data given an underlying process for yield, the process model that relates an underlying process (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 $k \in \{O, A, Q, G\}$ (for OYS, AYS, and quarterly APS and CG, respectively), in year $t \in \{1, 2, ..., T\}$ and month $m \in \{8, 9, 10, 11, 12, 13\}$. Conditional on the latent regional yield, μ_t, data models for forecast month m are described by

$$\begin{aligned} y_{ktm}|\mu_t &\sim \textit{indep } N\left(\mu_t + b_{km}, s^2_{ktm} + \sigma^2_{km}\right), \\ k &= O, A, Q, G;\ m < 13 \end{aligned} \tag{1}$$

$$y_{Gt,13}|\mu_t \sim \textit{ indep } N\left(\mu_t, s^2_{Gt,13}\right) \tag{2}$$

In this specification, observed survey yields and ginnings yield estimates are modeled with potential month-specific biases, whereas yields computed based on January ginnings are used as a proxy for the gold-standard May ginnings and assumed unbiased as shown in Equation 2. The region level process model varies around a mean based on a regression of historic end of season yield and observable covariates:

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

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

Under the assumption of conditional independence, the likelihood function has the multiplicative form

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

and by Bayes' Rule the posterior distribution of model parameters given observable yield estimates is shown in Equation 5:

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

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

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

where

$$\begin{aligned}
\Delta_1 &= \sum_{k=O,A} \frac{1}{\sigma^2_{km} + s^2_{ktm}} + \frac{I_{m\in\{10,\ldots,13\}}}{I_{\{m\neq 13\}}\sigma^2_{Gm} + s^2_{Gtm}} && (7)\\
&+ \frac{I_{\{m=13\}}}{\sigma^2_{Q,13} + s^2_{Qt,13}} + \frac{1}{\sigma^2_\eta} \\
\Delta_2 &= \sum_{k=O,A} \frac{y_{ktm} - b_{km}}{\sigma^2_{km} + s^2_{ktm}} + \frac{I_{m\in\{10,\ldots,13\}}(y_{Gtm} - I_{\{m\neq 13\}}b_{Gm})}{I_{\{m\neq 13\}}\sigma^2_{Gm} + s^2_{Gtm}} && (8)\\
&+ \frac{I_{\{m=13\}}(y_{Qt,13} - b_{Q,13})}{\sigma^2_{Q,13} + s^2_{Qt,13}} + \frac{\boldsymbol{z}_t'\boldsymbol{\beta}}{\sigma^2_\eta}.
\end{aligned}$$

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 6 is *shown to be a weighted average of available information sources*: the bias-corrected AYS and OYS indications, the bias corrected quarterly APS indication (when it is available), bias corrected ginnings in all months but January (when it is assumed unbiased), and covariates information. Since NASS does not publish the individual inputs, this relationship serves as a useful interpretation for the one number yield forecast as a *meaningful composite* of the available information based on posterior variance; the most precise information sources receive a proportionally larger share of weight in determining the overall yield forecast.

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:

$$\begin{aligned}
y_{ktmj}|\mu_{tj} &\sim indep\, N\left(\mu_{tj} + b_{kmj}, s^2_{ktmj} + \sigma^2_{kmj}\right), k = O, A, Q, G; m < 13 && (9)\\
y_{Gt,13,j}|\mu_{tj} &\sim indep\, N\left(\mu_{tj}, s^2_{Qtj}\right), && (10)\\
\mu_{tj} &\sim indep\, N\left(\boldsymbol{z}'_{tj}\boldsymbol{\beta}_j, \sigma^2_{\eta j}\right). && (11)
\end{aligned}$$

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 6. Assuming independence, the collection of state-level crop yields follows a multivariate normal distribution.

$$[\boldsymbol{\mu}_{t\cdot}|\boldsymbol{y}, \boldsymbol{\Theta}_d, \boldsymbol{\Theta}_p] \sim indep\, MVN\left(vec\left(\frac{\Delta_{2tj}}{\Delta_{1tj}}\right), diag\left(\frac{1}{\Delta_{1tj}}\right)\right) \tag{12}$$

While yield parameters for the region μ_t and states μ_{tj} must respect the balance identity $\mu_{tj} = \sum_j w_j \mu_{tj}$, estimates of parameters $\hat{\mu}_{tj}$ derived under Equation 12 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 12 on the speculative region yield μ_t, the collection of the first $j - 1$ states will follow a multivariate normal distribution

$$\left(\mu_{t1}, \mu_{t2}, \ldots, \mu_{t(J-1)}\right) \sim MVN(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\Sigma}}). \tag{13}$$

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

$$\mu_{tJ} = \mu_t - \frac{1}{w_{tJ}} \sum_{j=1}^{J-1} w_{tj} \mu_{tj}, \tag{14}$$

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 6, Equation 13, and Equation 14 represent a collection of point estimates for the speculative region and all its constituent states that honor the physical balance constraint. Standard errors of these estimates are derived as the square root of posterior variances, giving rise to defensible measures of uncertainty at both spatial scales.

4. A Candidate Model for 2016 Upland Cotton Yield

For illustrative purposes, we fit a model using the following process model for the j^{th} state:

$$\mu_{tj} \sim N\left(\beta_{j1} + \beta_{j2} PCP_j + \beta_{j3} CDD_{j3} + \beta_{j4} EXC_{j4}, \sigma_\eta^2\right) \tag{15}$$

where

- PCP is the state's average precipitation during the month of June
- CDD is the number of cooling degree days (a proxy for cumulative growing degree days) during June
- EXC is the percent of the cotton crop that has been rated excellent as of week 26 (late June to early July) according to NASS's crop condition ratings.

These reference dates were chosen based on correlation analysis with historical final cotton yields and the understanding that precipitation and growing degree days in June represent important requirements during a critical growth phase for cotton in most states.

For the specified process model, Figure 6 depicts the sequence of model-based yield forecasts from August through January during the 2016 crop year. For comparison, NASS official forecasts (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 candidate 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. For comparison, yields computed from May ginnings are represented by the single point.

A salient feature of the cotton yield forecasting 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 with weights proportional to a posteriori precision. For the 2016 crop year, the emphasis applied to each information source in the model-based forecasts may vary by month as shown in Figure 7. 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 and AYS estimates in October and November, and then to cotton ginnings and the quarterly APS survey in January.

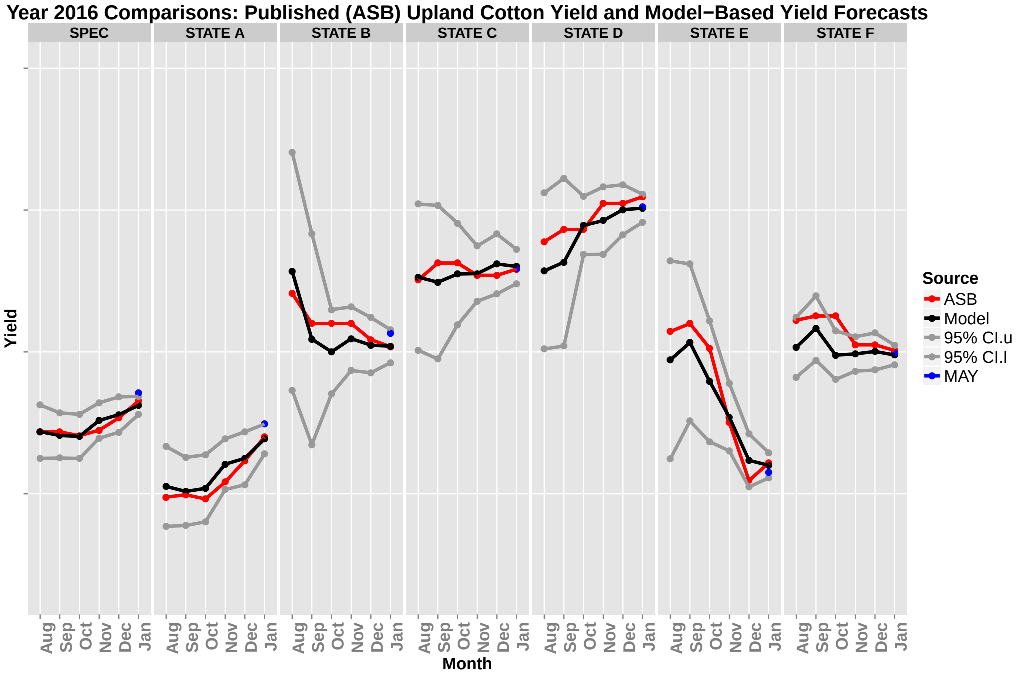


Figure 6: Sequence of model-based and official cotton yield forecasts for 2016 crop year

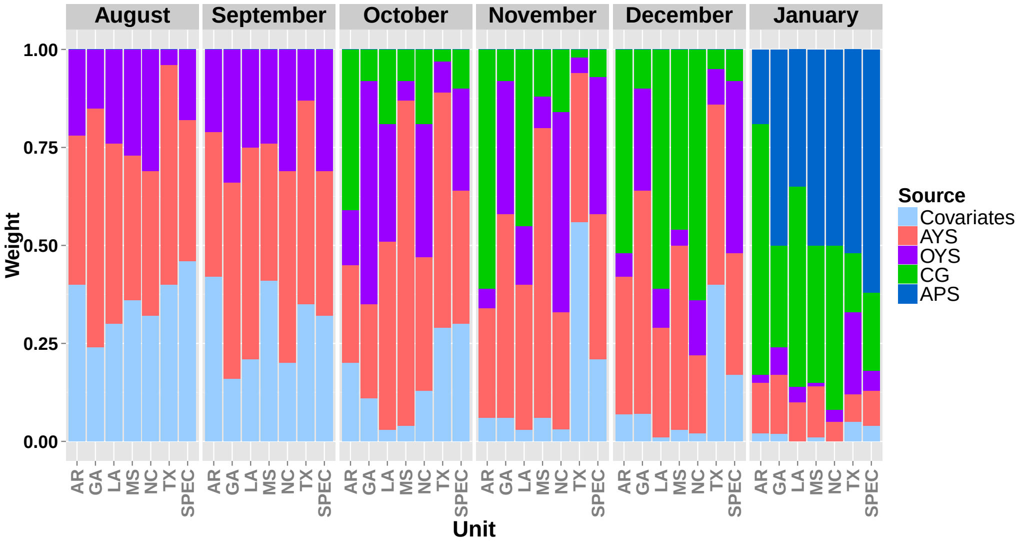


Figure 7: Relative emphasis of each input information source

5. Conclusions

Bayesian hierarchical models for crop yield have been used by NASS in recent years to provide additional useful estimates to ASB decision makers in support of official forecasts and estimates of corn, soybeans, and winter wheat. The existence of cotton ginnings data leads to a notion of 'gold standard' data source that is unique to upland cotton; measures of bias in early season survey estimates are derived relative to late season ginnings estimates. An estimated mean squared error was proposed as a means of quantifying the uncertainties in early-season cotton ginnings reports. The adaptations presented in this paper permit the use of the additional cotton ginnings projections and three probability-based NASS surveys to produce benchmarked forecasts of cotton yield at regional and state levels. Through inclusion in the process model, relevant weather data and crop condition ratings could also be incorporated. Ongoing research in variable selection for the process model could help improve the accuracy of early season model-based forecasts. As NASS seeks to demonstrate the ability to support the scope of commodities in its federally-mandated Crop Production Report through model-based techniques, extensions to the national program will require modeling yield in the absence of Objective Yield Survey estimates for non-speculative states. In the long run, modeling may provide a means of incorporating multiple, possibly disparate, estimates in a reproducible manner that gives rise to measures of uncertainty.

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.

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.

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.