Model-Based Crop Yield Forecasting: Adjustment for Within-State Heterogeneity, Covariate Selection and Variance Estimation

Habtamu K. Benecha[*], Nathan B. Cruze[*], Noemi Guindin[*], Nell Sedransk[†]

Abstract

The USDA's National Agricultural Statistics Service (NASS) produces monthly and annual yield forecasts for major crops at state and regional levels. To support the forecasting program, several monthly surveys are conducted during the growing season. In addition to the surveys, administrative data are also available for crops such as upland-cotton, for which production data are collected biweekly from cotton gins. The Research and Development Division at NASS has developed Bayesian hierarchical models that combine data from these surveys and several covariates, including weather data, to produce yield forecasts for each state where the crop is grown and for the region that comprises major crop producing states. The Bayesian approach is extended to adjust for heterogeneities in yield, production, weather and other factors within a state, by partitioning the state into more homogeneous sub-areas and then incorporating data from these sub-areas into the model. Alternative approaches to covariate selection are discussed; and estimation of measures of uncertainty associated with administrative data are considered. Performances of alternative models are compared.

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

1. Introduction

To fulfill its mission of providing timely, accurate and useful statistics in service of U.S. agriculture, USDA's National Agricultural Statistics Service (NASS) publishes hundreds of reports every year. One such publication, the Crop Production Report, is a monthly report released to the public in accordance with federal law. The report contains within-season *forecasts* of final production and harvested acreage totals, and yield per acre for major crops during the growing season. Another official report, the Crop Production Annual Summary, is published at the end of the growing season, and it contains preliminary final *estimates*. The official statistics in the Crop Production Report and the Crop Production Annual Summary are consensus estimates of the Agricultural Statistics Board (ASB), which is a panel of statisticians and commodity experts within NASS. Before the reports are published, members of the ASB meet in a secured location at the NASS headquarters and synthesize market-sensitive data from multiple surveys and auxiliary data to produce official estimates for relevant quantities at state, regional, and national levels. Thus, NASS has a vested interest in combining multiple sources of survey and non-survey data that become available as the events of the growing season are realized.

NASS researchers have produced Bayesian hierarchical models for crop yield forecasting in order to provide ASB decision makers with objectively forecasted crop yields with associated measures of uncertainty. These Bayesian hierarchical models were initially based on the pioneering work of Wang et al. (2012) and Nandram et al. (2014), and they have been refined for use in ASB process in support yield forecasts as described by Adrian

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

[†]National Institute of Statistical Sciences, 1750 K Street, NW, Suite 1100, Washington DC 20006-2306

(2012) (for corn and soybeans), Cruze (2015, 2016) (winter wheat), and Cruze and Benecha (2017) (upland cotton). The yield forecasting models facilitate multiple outcomes:

- combination of current and historical predictions of yield obtained from multiple surveys,
- incorporation of relavant auxiliary data and covariates such as weather information and crop condition,
- and resulting consistent one-number yield forecasts and measures of uncertainty for regions and member states.

Because the current yield models for upland cotton and the other crops input data at state and regional levels, any within-state heterogeneities in yield, production, and weather conditions may not be fully accounted for when forecasts are produced. For example, two distinct clusters of upland cotton grown in Texas, the largest cotton producing state in the U.S., differ in terms of climate and typical annual cotton yield.

In this paper, we extend the existing upland cotton yield model of Cruze and Benecha (2017) by incorporating sub-state level data for Eastern and Western districts in Texas, with the intention of improving the state and regional yield forecasts and the corresponding standard errors. The proposed model is compared with existing methods and its merits are assessed. Additionally, covariate selection and alternative approaches to characterizing the variability for non-probability cotton ginnings data are also discussed. The paper is organized as follows: In Section 2, the upland cotton speculative region and available sources of data for forecasting upland cotton yield in the context of the NASS publication timeline are described. In Section 3, yield estimates obtained from three distinct NASS surveys and a census of cotton processors (cotton gins) are compared. Estimates from two distinct clusters of cotton in Texas are compared. Section 4 describes a Bayesian hierarchical model for upland cotton. In Section 5, issues related to covariate selection and variances of yield estimates from cotton ginnings are discussed. In section 6, a preliminary analysis assessing the effects that result from assumptions about sub-state breakouts, choice of covariates, and assumptions of about uncertainty in cotton ginnings is presented. Concluding remarks are given in Section 7.

2. The speculative region, sources of data, and publication timelines

2.1 The upland cotton speculative region

Currently, NASS publishes estimates and forecasts of upland cotton yield, production, harvested acreage and related statistics every month from August through January for the nation and the 17 southern states shown in the map in Figure 1. The six states in the dark-shaded area constitute the current speculative region. Together they account for most of the upland cotton production in the nation. While the membership of the upland cotton speculative region has changed over the years, the six states (Arkansas, Georgia, Louisiana, Mississippi, North Carolina, and Texas) have constituted the region since 2008. Currently, the scope of the model-based approach is aimed at producing benchmarked and reproducible monthly yield forecasts and associated measures of uncertainty for these six states and the speculative region as a whole.

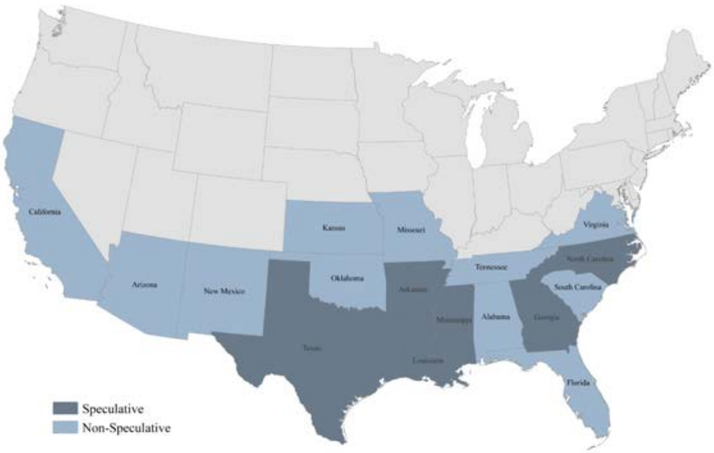

Figure 1: USDA NASS Upland Cotton Estimation Program States and Speculative Region

2.2 Sources of data for yield forecasting and NASS publication timelines

Forecasting and estimation of upland cotton yield in the Crop Production Report and the Crop Production Annual Summary is supported by a biweekly census of cotton gins in all cotton producing states, and by three probability-based surveys: the Objective Yield Survey (OYS), the Agricultural Yield Survey (AYS) and the December Quarterly Acreage, Production, and Stocks (APS) survey. Approximate data collection windows for each of these sources and the associated publication deadlines are depicted in Figure 2. The OYS is based on field measurements collected at sampled field plots. It is conducted monthly from August through January. The OYS covers only the six states in the speculative region, and it gives rise to monthly point predictions of regional and state yield with associated standard errors. The AYS is a monthly farmer interview survey conducted from August to November. Like the OYS, the AYS provides point predictions of state and regional level yield and standard error estimates. The third NASS survey, the December APS survey is a farmer interview survey conducted near the end of the growing season in December. The APS survey is conducted after much of the crop is harvested and involves larger sample sizes than the OYS and the AYS. As a result, the APS survey gives rise to more accurate estimates of yield and with lower sampling variation.

A fourth source of data for estimation of upland cotton statistics is derived from a biweekly census of cotton processing gins in cotton producing states. In this exhaustive census, cotton *processors* (note, not farmers) are requested to report:

1. the number of bales of cotton already processed in the season as of a specified reference date, and

2. the number of bales of cotton expected to be processed during the time interval from the reference date to the end of the crop year.

Based on these records, *total production* is projected for states and the nation. Although some states begin reporting ginnings data in earlier months, projected cotton ginnings production data are available for all producing states starting from October of each year. As a result, the ASB starts considering ginnings data in support of October forecasts, and continues to consider such data every month until the Crop Production Annual Summary report is released in January. By law, NASS must publish its official upland cotton yield and other statistics on or before the twelfth day of each month during the growing season. Approximate data collection windows for the three surveys and the first-of-month cotton ginnings are shown in Figure 2. The publication of August and September official statistics for upland cotton is mainly supported by the OYS and the AYS, and estimates from October through December are supported by data from OYS, AYS and cotton ginnings. The January estimates published in the Crops Annual Summary incorporate the December APS, cotton ginnings other survey data.

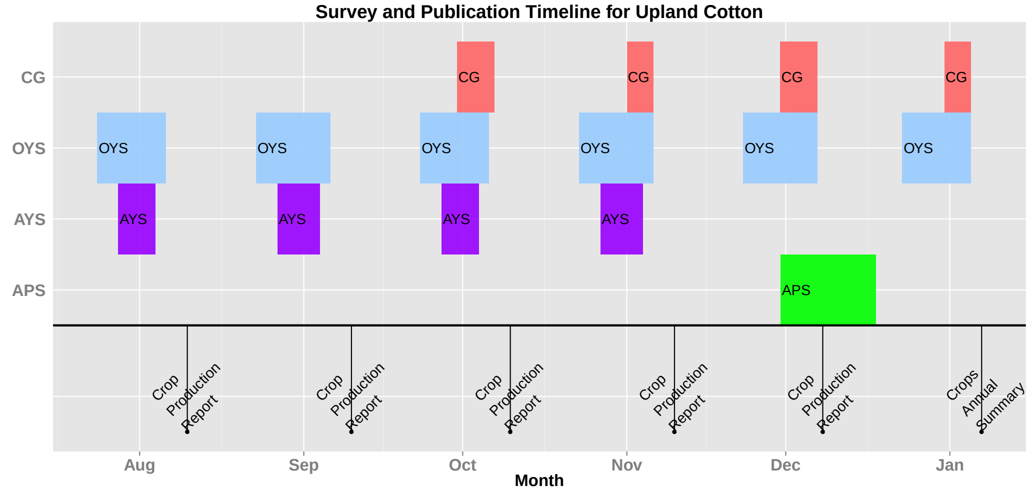

Figure 2: Survey and report production timeline for NASS upland cotton yield forecasts

While preliminary annual crop statistics for upland cotton are produced and reported in the Crop Production Annual Summary in January of the next calendar year, the census of cotton gins continues until May, as cotton processing in some states can evolve some months beyond January. Cotton growers are paid by the cotton gins for their cotton; thus, the total ginnings production in May represents a near-complete accounting of all upland cotton grown in the U.S. Thus, May cotton ginnings is thought of as a *gold standard* at state, regional, and national levels. Unlike the OYS, AYS and APS surveys; however, yield predictions (on the ratio scale) and corresponding sampling variances cannot be directly obtained from these biweekly censuses of cotton gins. Cruze and Benecha (2017) converted projected cotton production totals to the yield ratio scale by dividing by harvested area and referred to these derived point predictions as ginnings yields. *The nature of uncertainty in the ginnings yields has little to do with sampling error (all known gins are included) and more to do with the forecasting error in the early cotton ginnings projections* and any potential nonresponse by cotton gin operations. As a preliminary approach to assessing these uncertainties, Cruze and Benecha (2017) derived estimated mean squared errors as a proxy for 'sampling variances' by computing the squares of historical deviations of monthly ginnings yield from final May cotton yields and computing moving averages based on the previous 10 year history, which resulted in declining uncertainty in ginnings yield as the season progressed. One alternative to this approach is considered in Section 4.

3. Survey estimates and within-state heterogeneities in yield

The membership of the speculative region has changed over time (California was included as a seventh state prior to 2008), but survey data on yield, production, and harvested acreage are available for all six states in the current speculative region as far back as 2001. To allow for the use of longer historical data series in the Bayesian model, the speculative region was assumed to contain the same six states from 2001 to present and the regional yield, standard error, production and harvested acreage statistics were resummarized for the years from 2001-2007. Historically, yield estimates from the surveys and from monthly cotton ginnings show disparities among each other, and they tend to be biased even systematically away from the (gold standard) May final cotton yield estimates. Figure 3 shows a series of survey and ginnings yield estimates for an example state from 2001-2017. The OYS series (in yellow) tends to overstate yield relative to May (solid dark blue line). The AYS series (in green) tends to show systematic downward bias in comparison to final May yields. In this particular state, the December APS may also show a slight downward bias, although the magnitude is smaller than in the other two surveys.

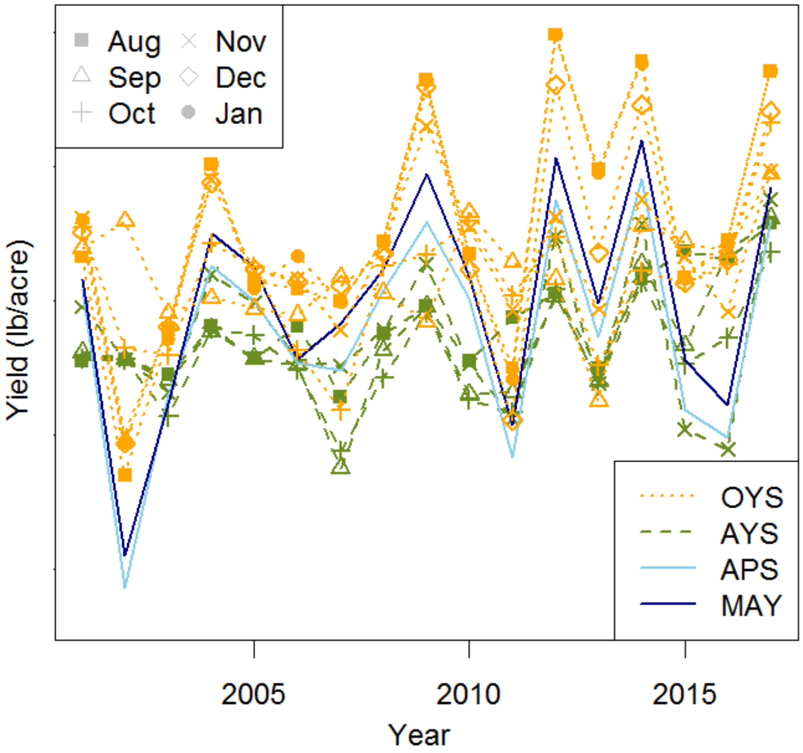

Figure 3: History of survey estimates relative to final May yields

In addition to the state and regional level estimates, NASS has traditionally produced OYS estimates for two sub-state regions in Texas for internal use. The two sub-states, which constitute the east (TE) and the west (TW) parts of Texas, have considerable differences in weather conditions, cotton yield and total production. The upland cotton annual harvested acreage from the west part of the state is much higher than that from the east part, but production per acre, i.e., yield, tends to be higher in the eastern districts than in the western part as shown in Figure 4. Because the existing cotton model inputs state-level estimates into the cotton model, within-state heterogeneities in yield and covariates (e.g., weather conditions) are not fully taken into account. In a preliminary effort to address this, AYS, APS, and ginnings, and covariates data from 2001 to present were resummarized for Texas based on the TE and TW breakouts. Each sub-state of Texas by itself accounts for larger harvested acreage, production and sample size than many of the other states in the speculative region and may be treated in the model as a ‘state’. Utilization of sub-state level data in the model fitting processes may result in improved yield forecasts for the state of Texas and for the speculative region as a whole. In addition to fitting models for the six states, we investigate a scenario in which the two regions of Texas enter the models as two separate states, the speculative region constitutes seven 'states' (i.e., AR, GA, LA, MS, NC, TE and TW) and the regional survey and cotton ginnings estimates are re-calculated by excluding the state level TX data and by including estimates from TE, TW.

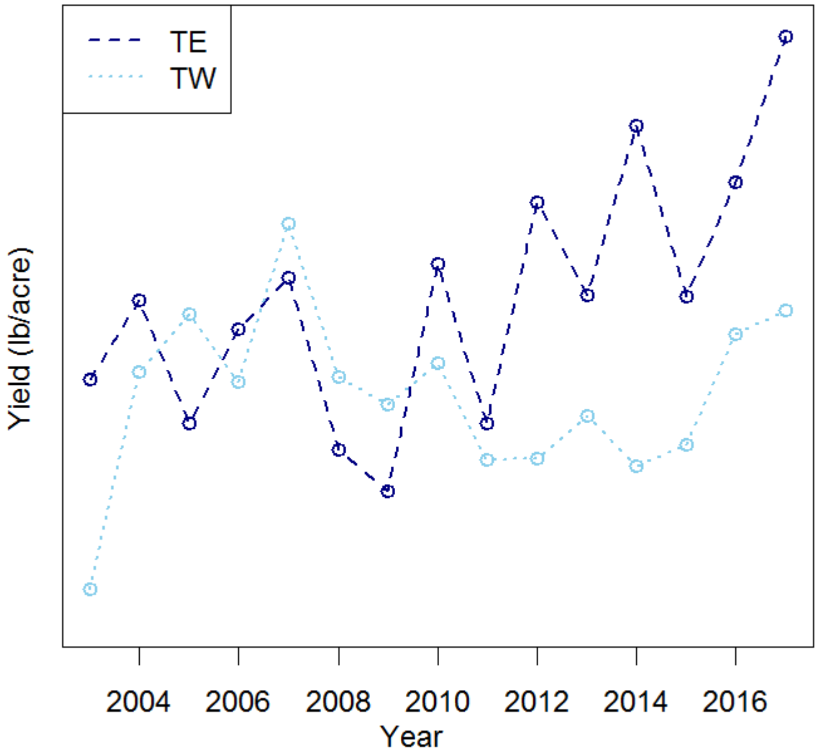

Figure 4: Annual yield estimates for two sub-states

4. Bayesian hierarchical model for yield forecasting

In this section, we outline extensions of the upland cotton yield model described in Cruze and Benecha (2017). The extended model allows for the estimation of parameters at state, regional and sub-state levels for the two regions of Texas. Whereas Cruze and Benecha (2017) initially treated January ginnings yield as a 'gold standard' (in keeping with the publication window), we introduce the May final yield as the gold standard and the quantity to be forecasted. For Texas in particular, where processing activity evolves beyond the Annual Summary publication window, this decision represents a better choice of gold standard.

4.1 Models for the speculative region

As discussed in Cruze and Benecha (2017), the Bayesian hierarchical models for the speculative region and its member states specify conditional and marginal distributions for the data and the parameters in three parts. The behavior of observed data given an underlying process for yield is described in a data model and the parameter of interest (i.e., yield, denoted by μ_t for the speculative region) is related to covariates of interest through a process model and prior distributions are specified for model parameters. Let y_{ktm} denote observed yield estimates from data source $k \in \{O, A, Q, G, M\}$ for OYS, AYS, APS, ginnings yield (from October to January), and the May final yield respectively, in year $t \in \{1, 2, ..., T\}$ and month $m \in \{8, 9, 10, 11, 12, 13\}$. Let s^2_{ktm} denote the variance of the yield estimate from source k in year t and month m. Assume for now that a variance estimate, s^2_{Gtm}, is available for ginnings yield. Conditional on the latent regional yield, μ_t, data models for forecast month m are described by

$$\begin{aligned}
y_{ktm}|\mu_t &\sim \quad indep\ N\left(\mu_t + b_{km}, s^2_{ktm} + \sigma^2_{km}\right), k = O, A,\ m \leq 13 && (1)\\
y_{Qtm}|\mu_t &\sim \quad indep\ N\left(\mu_t + b_{Qm}, s^2_{Qtm} + \sigma^2_{Qm}\right), m = 13 && (2)\\
y_{Gtm}|\mu_t &\sim \quad indep\ N\left(\mu_t + b_{Gm}, s^2_{Gtm} + \sigma^2_{Gtm}\right), m = 10, 11, 12, 13 && (3)\\
y_M|\mu_t &\sim \quad indep\ N\left(\mu_t, \sigma^2_M\right) && (4)
\end{aligned}$$

In this specification, observed survey yields and ginnings yield estimates are modeled with potential month-specific biases, whereas the May final yield estimates are used as a proxy for the gold-standard May ginnings and assumed unbiased as shown in Equation 4. Although the last AYS survey of the season is conducted in November, estimates from the November survey may be included in the analyses for making the December and January forecasts. Note also that estimates from the December Quarterly APS survey are used in the January final model; data collection for the APS is onoing when December forecasts are due for publication. Given an appropriate sampling variance estimate, s^2_{Gtm}, the data model for yield from monthly ginnings takes the form shown in Equation 3. We consider two scenarios, one in which an estimate of s^2_{Gtm} is obtained separately, and another in which we assume $\hat{s}^2_{Gtm} = 0$.

The region-level process model varies around a mean based on a regression of historic end-of-season yield on observable covariates:

$$\mu_t \sim \ indep\ N\left(\boldsymbol{z}_t'\boldsymbol{\beta}, \sigma^2_\eta\right). \qquad (5)$$

Finally, the following prior distributions are specified for the parameters; b_{km}, $\boldsymbol{\beta} \sim$ indep $N(0, 10^6)$, $\sigma^2_{km}, \sigma^2_\eta, \sigma^2_{Gtm} \sim$ indep $IG(.001, .001)$, and $\sigma^2_M \sim$ indep Uniform $(.0005, .001)$. The collection of data and process model parameters are denoted $\boldsymbol{\Theta}_d \equiv \left(b_{km}, \sigma^2_{km}, \gamma^2_{Gm}, \sigma^2_M\right)$ 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

$$[y_O, y_A, y_Q, y_G|\mu_t, \boldsymbol{\Theta}_d] = \prod_{k \in \{O,A,Q,G\}} [y_k|\mu_t, \boldsymbol{\Theta}_d] \qquad (6)$$

and by Bayes' Rule, the posterior distribution of model parameters given observable yield estimates is shown in Equation 7:

$$[\mu_t, \boldsymbol{\Theta}_d, \boldsymbol{\Theta}_p|y_O, y_A, y_Q, y_G] \propto \prod_{k \in \{O,A,Q,G\}} [y_k|\mu_t, \boldsymbol{\Theta}_d][\mu|\boldsymbol{\Theta}_p][\boldsymbol{\Theta}_d][\boldsymbol{\Theta}_p]. \qquad (7)$$

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:

$$[\mu_t|y_O, y_A, y_Q, y_G|\boldsymbol{\Theta}_d, \boldsymbol{\Theta}_p] \sim N\left(\frac{\Delta_2}{\Delta_1}, \frac{1}{\Delta_1}\right) \qquad (8)$$

where

$$\Delta_1 = \sum_{k=O,A} \frac{1}{\sigma^2_{km} + s^2_{ktm}} + \frac{I_{m\in\{10,\ldots,13\}}}{\sigma^2_{Gtm} + s^2_{Gtm}} + \frac{I_{\{m=13\}}}{\sigma^2_{Q,13} + s^2_{Qt,13}} + \frac{1}{\sigma^2_\eta} \tag{9}$$

$$\Delta_2 = \sum_{k=O,A} \frac{y_{ktm} - b_{km}}{\sigma^2_{km} + s^2_{ktm}} + I_{m\in\{10,\ldots,13\}} \frac{y_{Gtm} - b_{Gm}}{\sigma^2_{Gtm} + s^2_{Gtm}} \tag{10}$$

$$+ \quad \frac{I_{\{m=13\}}(y_{Qt,13} - b_{Q,13})}{\sigma^2_{Q,13} + s^2_{Qt,13}} + \frac{\boldsymbol{z}_t'\boldsymbol{\beta}}{\sigma^2_\eta}.$$

Equation 9 describes the sum of the precisions of each information source. Dividing Equation 10 by Equation 9, the mean of the full conditional distribution Equation 8 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 , 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.

4.2 Models for states

We consider two alternative models for states: one for the six states in the current speculative region (i.e, AR, GA, LA, MS, NC and TX), and one in which the two sub-state regions of Texas enter the model as separate 'states', i.e., AR, GA, LA, MS, NC, TE, and TW. and the regional survey, ginnings and covariate data are re-calculated based on the new set of member states. In the following discussions, the state index j takes values from 1 to 6 when TX is included in the speculative region and takes values from 1 to 7 when TE and TW are included in the region as two distinct states. Data and process models for the states resemble those of the speculative region with models for each state j given by:

$$y_{ktmj}|\mu_t \sim indep\, N\left(\mu_{tj} + b_{kmj}, s^2_{ktmj} + \sigma^2_{kmj}\right), k = O, A,\ m \leq 13 \tag{11}$$

$$y_{Qtmj}|\mu_t \sim indep\, N\left(\mu_{tj} + b_{Qmj}, s^2_{Qtmj} + \sigma^2_{Qmj}\right), m = 13 \tag{12}$$

$$y_{Gtmj}|\mu_t \sim indep\, N\left(\mu_{tj} + b_{Gmj}, s^2_{Gtmj} + \sigma^2_{Gtmj}\right), m = 10, 11, 12, 13 \tag{13}$$

$$y_{Mj}|\mu_t \sim indep\, N\left(\mu_{tj}, \sigma^2_{Mj}\right) \tag{14}$$

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

$$[\boldsymbol{\mu}_{t\cdot}|\boldsymbol{y}, \boldsymbol{\Theta}_d, \boldsymbol{\Theta}_p] \sim indep\, MVN\left(vec\left(\frac{\Delta_{2tj}}{\Delta_{1tj}}\right), diag\left(\frac{1}{\Delta_{1tj}}\right)\right) \tag{15}$$

While yield parameters for the region μ_t and states μ_{tj} must respect the balance identity $\mu_t = \sum_j w_j \mu_{tj}$, estimates of parameters $\hat{\mu}_{tj}$ derived under Equation 15 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 15 on the restriction that their weighted sum is equal to forecasted speculative region yield μ_t, the collection of the first $j-1$ states will follow a multivariate normal distribution

$$\left(\mu_{t1}, \mu_{t2}, \ldots, \mu_{t(J-1)}\right) \sim MVN(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\Sigma}}). \tag{16}$$

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

$$\mu_{tJ} = \mu_t - \frac{1}{w_{tJ}} \sum_{j=1}^{J-1} w_{tj} \mu_{tj}, \tag{17}$$

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 8, Equation 16, and Equation 17 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.

5. Covariates and variances for ginnings yield

5.1 Covariates and covariate selection

Estimates of the latent mean parameters (i.e., μ_t and μ_{tj}) in Sections 4.1 and 4.2 are related to a number of factors that affect upland cotton yield. The existing cotton model includes average precipitation (PCP), average cooling degree days (CDD), crop condition ratings (COND) and a drought sevierity index (DRT) as covariates to model μ_t and μ_{tj}. These covariates were choosen based on a combination of exploratory analysis and expert suggestions. Using a similar approach, two more predictors are added to the pool of potential covariates for consideration in our analysis. The additional covariates are monthly maximum temprature (TMP) and monthly killing degree days (KDD), resulting in a total of six potential covariates (i.e., PCP, CDD, COND, DRT, TMP and KDD) to choose from. An important task in selecting predictors from the pool is determining the week or month in which each covariate has the highest impact on yield. Determining a subset of the pool of covariates that gives an optimal fit to the model is another important consideration. Preliminary attempts were made to address these issues by applying formal covariate selection procedures. For example, spike-and-slab priors, as described in Ishwaran and Rao (2005), were applied to select the optimal covariate combination with the month in which each selected covariate impacts yield the most. Stepwise regression and least absolute shrinkage and selection operator (LASSO) approaches were also applied to select covariates for a model that uses the May final yield as a response variable. Application of the variable selection procedures resulted in different sets of optimal covariates for different states. In addition, selected covariate sets vary by forecasting months in a season. At this time, we consider a common set of covariates observed on each state to make yield forecasts.

The covariate selection process is still an ongoing research project. For comparisons, model-based forecasts were made using a modified version of the covariates in the existing model (i.e., $\{$PCP, CDD, COND, DRT$\}$) and by using a tentatively selected set of covariates $\{$PCP, TMP, KDD, COND$\}$. In year t, the distribution of the latent mean yield parameter for state i can be expressed using the two covariate sets as

$$\begin{aligned}
\mu_{tj} &\sim N\left(\beta_{j1} + \beta_{j2} PCP_j + \beta_{j3} CDD_j + \beta_{j4} COND_j + \beta_{j5} DRT_j, \sigma_\eta^2\right), \text{and} \\
\mu_{tj} &\sim N\left(\beta_{j1} + \beta_{j2} PCP_j + \beta_{j3} TMP_j + \beta_{j4} KDD_j + \beta_{j5} COND_j, \sigma_\eta^2\right).
\end{aligned}$$

Where,

- CDD$_j$ is the number of cooling degree days (a proxy for cumulative growing degree days) during July
- PCP$_j$ is the state's average precipitation during the month of July
- COND$_j$: is the percent of the cotton crop that has been rated excellent as of week 28 according to NASS's crop condition ratings.
- DRT$_j$ is percent of land in the sate with extreme and exceptional drought in July
- KDD$_j$ is average July killing degree days
- TMP$_j$ Maximum temprature during July.

For the speculative region model, the corresponding covariate values are calculated as weighted sums of the state level covariates.

5.2 Variances for ginnings yield

Unlike the OYS, AYS and APS surveys, yield estimates from ginnings have no associated design-based variance estimates. As described in Sections 4.1 and 4.2, the model assumes that the variances of the conditional distributions of yield estimates from OYS, AYS and APS are sums of the corresponding sampling variances and a parameter characterizing variation due to 'other' non-sampling sources as in

$$y_{Gtmj}|\mu_t \quad \sim \quad indep\ N\left(\mu_{tj} + b_{Gmj}, s^2_{Gtmj} + \sigma^2_{Gtmj}\right), m = 10, 11, 12, 13. \quad (18)$$

To allow for a similar variance specification for ginnings yield estimates, Cruze and Benecha (2017) initially estimated variances for ginnings yield using deviations of the monthly ginnings yield estimates from the May final yield. In some since, this approach may have penalized ginnings data twice, since the biases in early season ginnings drove both the s^2_{Gtmj} and the term related to non-sampling errors. In this work, we consider two alternative approaches:

- Assumption 1: $\hat{s}^2_{Gtmj} = 0$,
- Assumption 2: $\hat{s}^2_{Gtmj}$ is an estimated quantity.

Under the second assumption, we estimate s^2_{Gtmj} based on a model for the May final yield that includes monthly ginnings yield and OYS yield as covariates.

6. Application to the 2015, 2016 and 2017 upland cotton yield

The model and the methods discussed in Section 4 were applied to revisit the 2015, 2016 and 2017 upland cotton yield forecasts. Models were fitted for the current speculative region and its member states (i.e., AR, GA, LA, MS, NC, and TX). Another set of models were fitted for AR, GA, LA, MS, NC, TE and TW, and the speculative region, based on re-calculated survey, ginnings and covariate estimates. In the following discussions, we refer to the first model as the six-state model and the second model as the seven-state model. Each of the two models were fitted using two different sets of covariates (i.e., (PCP, CDD, COND, DRT) and (PCP, TMP, KDD, COND)) and under the two assumptions about the variances for the conditional distributions of ginnings yield estimates from October to January. In all, these represent $2 \times 2 \times 2$ specifications. We describe select outcomes below.

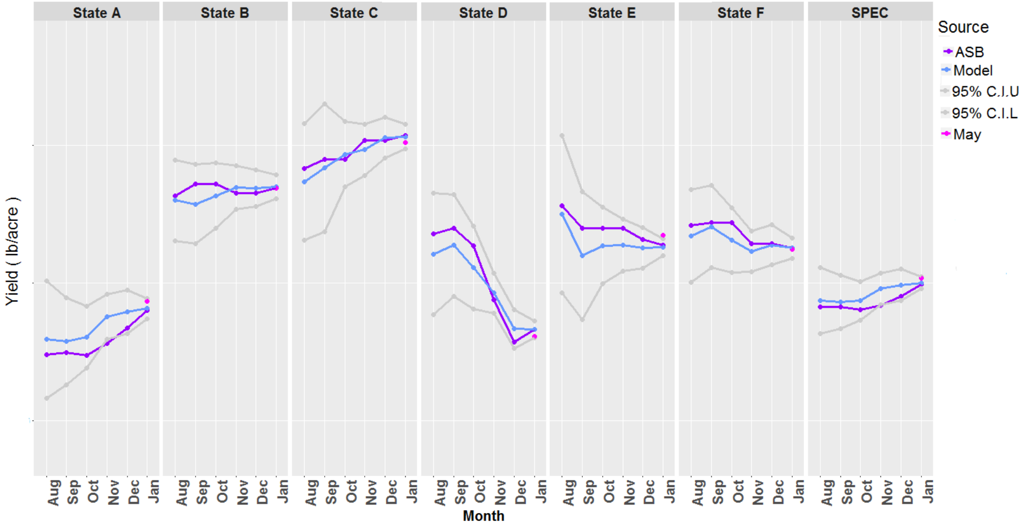

Figure 5: Published and model forecasts based on covariate set {PCP,CDD,DRT,COND} for 2016

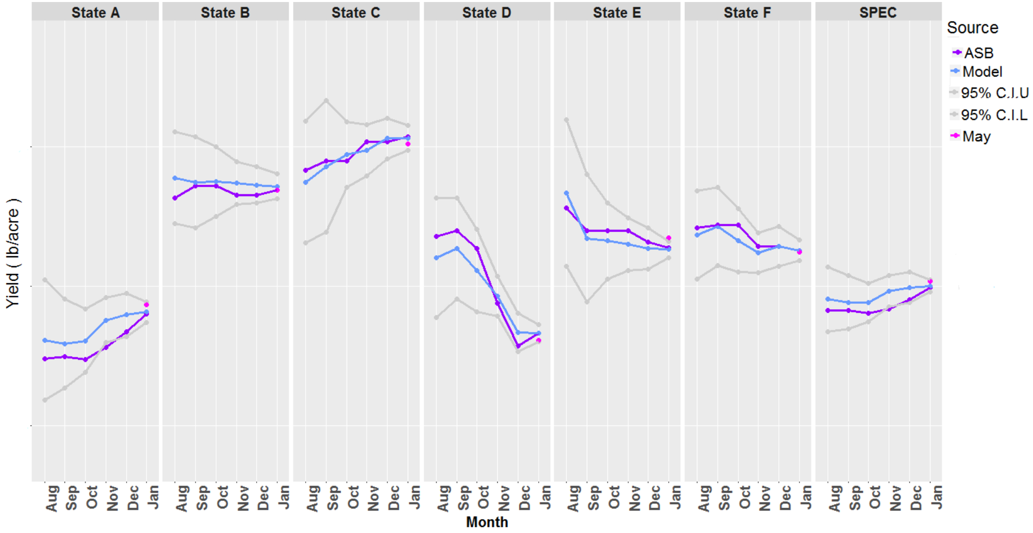

Figure 6: Published and model forecasts based on covariate set {PCP,TMP,KDD,COND} for 2016

Figures 5 and 6 show the 2016 forecasts based on covariate sets (PCP, CDD, COND, DRT) and (PCP, TMP, KDD, COND) applied to the seven state model with Assumption 2 for the variance of the conditional distribution of ginnings yield. States A, B, C, D, E and F represent the six member states of the speculative region, where the model-based forecasts for Texas shown in the figures are calculated as a weighted average of the forecasts for TE and TW. The actual numbers and the state names are not shown because of disclosure limitations. The two figures and similar analyses for 2015 and 2017 (not shown) show that forecasts based on the two covariate sets are very close to each other for the last few months of the forecasting season and that the two covariate sets provide similar forecasts at the end of the season. The major difference in the forecasts from the two sets of covariates is in the early season forecasts. Generally, the effects of covariates on yield forecasts decrease monthly from August to January.

Table 1: Average percent absolute differences of model estimates from May yield

State	6-State Model	7-State Model
August Estimates		
A	9.4	10.1
B	2.5	3.6
C	8.9	9.8
D	22.3	24.0
E	6.1	8.5
F	7.6	8.6
$SPEC$	5.4	5.7
January Estimates		
A	1.2	1.2
B	0.9	0.7
C	1.7	2.1
D	2.1	2.1
E	2.7	2.8
F	0.4	0.7
$SPEC$	0.8	0.8

Figures 7 and 8 show forecasts from the six and the seven-state models for 2016 based on the covariate set (PCP, CDD, COND, DRT) and under Assumption 2 for the variance parameter σ^2_{Gtmj}. Estimates from the two models are generally close to each other, especially at the end of the season. To compare the two models, we computed the average percent absolute differences of monthly forecasts from the May final yield using the 2015, 2016 and 2017 model-based forecasts. Table 1 shows the average percent absolute differences for the August and the January forecasts.

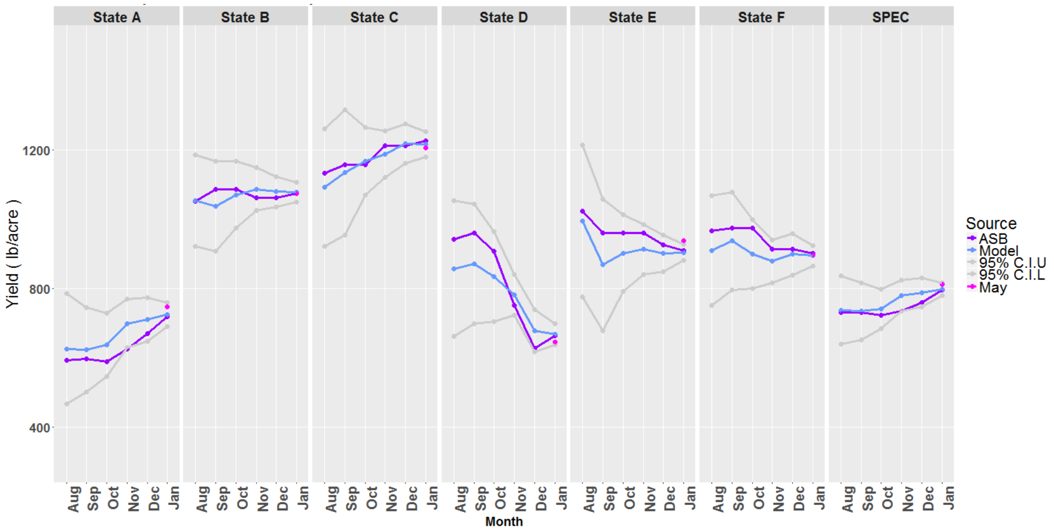

Figure 7: Published and model forecasts based on the six-state model

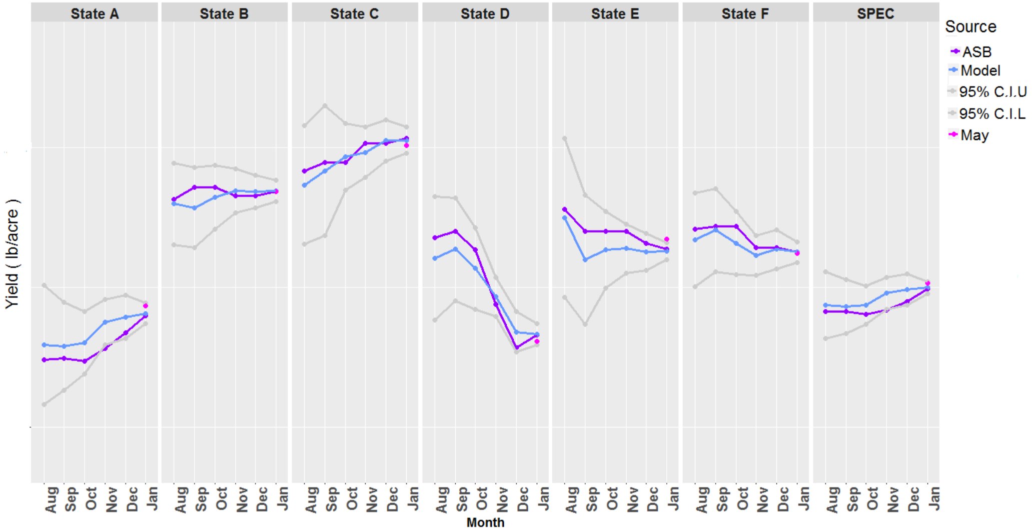

Figure 8: Published and model forecasts based on the seven-state model

The absolute differences from the two models are close to each other, but the six-state model performed better for most states. Finally, the two assumptions about the variance for the conditional distribution of ginnings yield were compared using the 2015, 2016 and 2017 model-based forecasts and the corresponding standard errors for the six-state model with covariate set (PCP, CDD, COND, DRT).

Table 2: Ratio of model estimated SEs for yield forecasts from models that use and do not use ginnings SEs

State	2015	2016	2017
August Estimates			
A	1.01	1.00	1.02
B	0.94	0.97	0.96
C	0.98	1.00	1.00
D	0.94	0.97	0.98
E	1.10	0.99	0.99
F	0.97	0.95	0.96
SPEC	1.01	0.99	1.00

Generally, the two assumptions provided similar forecasts and standard errors for all states and the speculative region, implying that we may not need to estimate sampling variances for ginnings yield separately. Ratios of model estimated SEs from the model under Assumptions 1 and 2 are presented in Table 2.

7. Discussion

A Bayesian hierarchical model that combines historical and current data from multiple surveys, cotton ginnings and covariates to produce a single forecast for a region and its member states has been developed. The approach allows for the estimation of a reproducible yield forecast with an associated measure of uncertainty. Because the current yield models for upland cotton and the other crops input data at state and regional levels, any within-state heterogeneities in yield, production, and weather conditions may not be fully accounted for when forecasts are produced. In an attempt to account for some heterogeneities and improve yield forecasts, sub-state level data were incorporated in the model-based analysis for the largest producing state, and comparisons were made between models that use state level data and models based on both state and sub-state level data. Comparisons of two assumptions for the variance of ginnings yield in the models showed little difference between the two approaches, indicating that the model may be fitted without the need for inputing an estimated variance for ginnings yield from October to January. Upland cotton yield is affected by a number of factors that can be incorporated into the model as covariates. As many of these covariates are related to weather conditions, an important task in selecting predictors from a pool is determination of the week or month in which a potential covariate has the highest impact on yield. In addition, determining the set of covariates that gives an optimal fit to the model is another important consideration. Preliminary attempts were made to select the best set of covariates for the upland cotton model, but covariate selection is still an ongoing research project.

Acknowledgements

The authors would like to thank Jared Pratt and several other USDA NASS Methodology Division staff for their knowledge and assitance in resummarizing an extended history of survey data and data for sub-state divisions of Texas. We thank Clyde Fraisse, and Ana Wagner of Univeristy of Florida for early access to smaller-than-state summaries of climatalogical variables derived from the AgroClimate tool. This research was supported in part by the intramural research program of the U.S. Department of Agriculture, National Agricultural Statistics Service.

Disclaimer

The findings and conclusions in this preliminary publication have not been formally disseminated by the U.S. Department of Agriculture and should not be construed to represent any agency determination or policy.

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.

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.

Cruze, N. B. and Benecha, H. K. (2017). A Model-Based Approach to Crop Yield Forecasting. In JSM Proceedings, Bayesian Statistical Science Section. Alexandria, VA: American Statistical Association.

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

Ishwaran, H. and Rao, J. S. (2005). Spike and slab variable selection: frequentist and Bayesian strategies. *The Annals of Statistics*, 2:730773.

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.