Model-Based Crop Yield Forecasting: Covariate Selection and Related Issues

Habtamu K. Benecha*, Luca Sartore*†, Nathan B. Cruze*

Abstract

USDA's National Agricultural Statistics Service (NASS) publishes hundreds of reports every year. Such publications include monthly and annual yield forecasts and estimates for major crops. To produce the forecasts, several surveys are conducted during the growing season. In recent years, NASS has been applying Bayesian hierarchical models to combine summaries from multiple surveys, administrative data and several covariates to produce a single estimate for a state or a region that comprises major crop producing states; the model estimates supplement NASS's yield forecasting program. The influences of covariates on forecasted yield generally decrease through the growing season, but model covariates play an important role in early season forecasting. Currently, covariates being considered in the models are selected based on expert knowledge of crop development and growth dynamics. In this paper, formal variable-selection approaches are considered for the identification of optimal covariates. The best sets of covariates are then compared using differences between model-based early-season forecasts and final official annual yield estimates.

Key Words: Variable selection;Yield forecasting; Composite estimation; Survey sampling; Bayesian hierarchical model

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 publishes hundreds of reports every year. Such publications include the Crop Production Report, which is a monthly report released to the public in accordance with federal law. The report contains within-season *forecasts* of final production, 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 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 secure 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 developed Bayesian hierarchical models for crop yield forecasting in order to provide ASB decision makers with objective crop yield forecasts with associated measures of uncertainty. These models refine the pioneering works of Wang et al. (2012) and Nandram et al. (2014) for use in ASB processes in support of yield forecasts as described by Adrian (2012) (for corn and soybeans), Cruze (2015, 2016) (winter wheat), and Cruze and Benecha (2017) and Benecha et al. (2018) (upland cotton). The yield forecasting models combine current and historical predictions of yield obtained from

*USDA National Agricultural Statistics Service (NASS), Room 6409A–South Building, 1400 Independence Ave., SW, Washington, DC 20250

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

multiple surveys, relavant auxiliary data and covariates to produce consistent one-number yield forecasts and measures of uncertainty for regions and member states.

Currently, covariates included in the models are selected based on expert knowledge of crop development and growth dynamics. This paper is focused on applying variable selection approaches to identify optimal covariate sets for the models. Because the NASS yield models for the different crops are similar, we focus our discussions on the upland cotton yield forecasting model. In Section 2, the upland cotton speculative region, its member states and available sources of data for forecasting yield in the context of the NASS publication timeline are described. Section 3 describes the Bayesian hierarchical model for upland cotton. In Section 4, model covariates, pool of potential covariates, dimension reduction, covariate selection, and forecasting performances of several covariate sets are discussed. Concluding remarks are given in Section 5.

2. The speculative region and data sources

2.1 The speculative region and its member states

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 four states with the darker coloration constitute the speculative region as of the 2019 crop season. These are the top upland cotton producing states in the nation, accounting for at least 65% of total production in the nation over the last five years. Membership of the upland cotton speculative region has changed over the years; from 2008-2018 there were six states in the region and starting from this year four states (Arkansas, Georgia, Mississippi and Texas) make up the region. Currently, the scope of the model-based approach aims at producing benchmarked and reproducible monthly yield forecasts and associated measures of uncertainty for these four states and the speculative region as a whole.

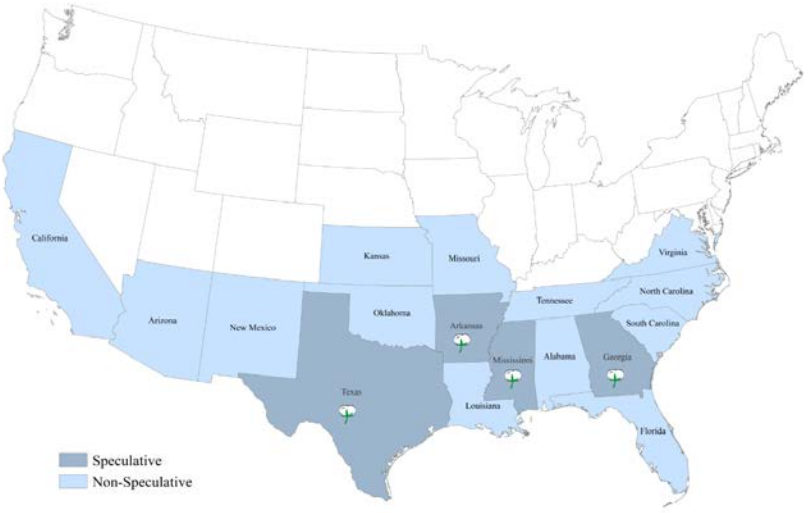

Figure 1: USDA NASS Upland Cotton Estimation Program States and Speculative Region

2.2 Sources of data for yield forecasting and NASS publication timelines

The upland cotton yield forecasting and estimation program 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 shown in Figure 2. The OYS is based on field measurements collected at sampled field plots. It is conducted monthly from September through January. The OYS covers only the four 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.

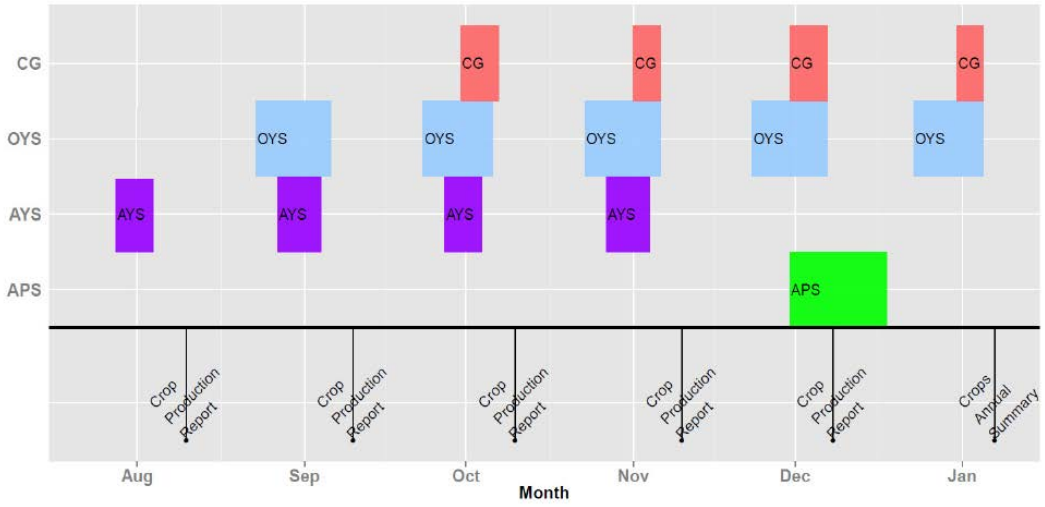

Figure 2: Survey and report production timeline for NASS upland cotton yield forecasts

A fourth source of data for producing upland cotton statistics is derived from a bi-weekly 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* for states and the nation is projected. 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.

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 survey estimates, however, yield predictions (on the ratio scale) and corresponding sampling variances cannot be directly obtained from these biweekly censuses of cotton gins.

3. Bayesian Hierarchical Model for Yield Forecasting

3.1 Models for the speculative region

The Bayesian hierarchical crop yield forecasting models refine the works of Wang et al. (2012) and Nandram et al. (2014) for use in the ASB process in support of yield forecasts for corn and soybeans (Adrian, 2012), winter wheat (Cruze, 2015; Cruze, 2016), and upland cotton (Cruze and Benecha, 2017; Benecha et al., 2018).

The upland cotton 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, 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 (October-January), and May final yield, respectively), in year $t \in \{1, 2, ..., T\}$ and month $m \in \{8, 9, 10, 11, 12, 13\}$, where $m = 13$ represents January. Let s^2_{ktm} denote the variance of the yield estimate from source $k \in \{O, A, Q\}$ in year t and month m. For the speculative region, conditional on the latent regional yield, μ_t, data models for forecast month m are described by

$$y_{ktm}|\mu_t \overset{\text{ind}}{\sim} \mathsf{N}\left(\mu_t + b_{km}, s^2_{ktm} + \sigma^2_{km}\right), k = A,\ m = 8, \tag{1}$$

$$y_{ktm}|\mu_t \overset{\text{ind}}{\sim} \mathsf{N}\left(\mu_t + b_{km}, s^2_{ktm} + \sigma^2_{km}\right), k = O, A,\ 8 < m \leq 13, \tag{2}$$

$$y_{Qtm}|\mu_t \overset{\text{ind}}{\sim} \mathsf{N}\left(\mu_t + b_{Qm}, s^2_{Qtm} + \sigma^2_{Qm}\right), m = 13, \tag{3}$$

$$y_{Gtm}|\mu_t \overset{\text{ind}}{\sim} \mathsf{N}\left(\mu_t + b_{Gm}, \sigma^2_{Gtm}\right), m = 10, 11, 12, 13, \tag{4}$$

$$y_M|\mu_t \overset{\text{ind}}{\sim} \mathsf{N}\left(\mu_t, \sigma^2_M\right). \tag{5}$$

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. 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.

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

$$\mu_t \overset{\text{ind}}{\sim} \mathsf{N}\left(\boldsymbol{z}_t'\boldsymbol{\beta}, \sigma^2_\eta\right). \tag{6}$$

Finally, vague, proper prior distributions complete the specification of model; for b_{km} and $\boldsymbol{\beta} \overset{\text{ind}}{\sim} \mathsf{N}(0, 10^6)$, and σ^2_{km}, σ^2_η, and $\sigma^2_{Gtm} \overset{\text{ind}}{\sim} \mathsf{IG}(.001, .001)$. A prior for σ^2_M is specified as $\sigma^2_M \overset{\text{ind}}{\sim} \mathsf{Uniform}(.0005, .001)$. 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

$$[y_O, y_A, y_Q, y_G, y_M|\mu_t, \boldsymbol{\Theta}_d] = \prod_{k\in\{O,A,Q,G,M\}} [y_k|\mu_t, \boldsymbol{\Theta}_d] \tag{7}$$

and based on Bayes' Rule, the posterior distribution of model parameters given observable yield estimates is:

$$[\mu_t, \boldsymbol{\Theta}_d, \boldsymbol{\Theta}_p|y_O, y_A, y_Q, y_G, y_M] \propto \prod_{k\in\{O,A,Q,G,M\}} [y_k|\mu_t, \boldsymbol{\Theta}_d][\mu|\boldsymbol{\Theta}_p][\boldsymbol{\Theta}_d][\boldsymbol{\Theta}_p]. \tag{8}$$

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:

$$[\mu_t|y_O, y_A, y_Q, y_G, y_M, \boldsymbol{\Theta}_d, \boldsymbol{\Theta}_p] \sim \mathsf{N}\left(\frac{\Delta_2}{\Delta_1}, \frac{1}{\Delta_1}\right) \tag{9}$$

where,

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

$$\Delta_2 = \sum_{k=O,A} \frac{y_{ktm} - b_{km}}{\sigma^2_{km} + s^2_{ktm}} + I_{m\in\{10,\ldots,13\}} \frac{y_{Gtm} - b_{Gtm}}{\sigma^2_{Gm}} + + \frac{I_{\{m=13\}}(y_{Qt,13} - b_{Q,13})}{\sigma^2_{Q,13} + s^2_{Qt,13}} + \frac{\boldsymbol{z}'_t\boldsymbol{\beta}}{\sigma^2_\eta}. \tag{11}$$

Equation 10 describes the sum of the precisions of each information source. Dividing Equation 11 by Equation 10, the mean of the full conditional distribution Equation 9 is *shown to be a weighted average of available sources of information*: 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.

3.2 Models for states

Data and process models for the states resemble those of the speculative region, with the model for state $j \in \{\text{AR, GA, MS, TX}\}$ given by:

$$y_{ktmj}|\mu_{tj} \overset{\text{ind}}{\sim} \mathsf{N}\left(\mu_{tj} + b_{kmj}, s^2_{ktmj} + \sigma^2_{kmj}\right), k = A,\ m = 8 \tag{12}$$

$$y_{ktmj}|\mu_{tj} \overset{\text{ind}}{\sim} \mathsf{N}\left(\mu_{tj} + b_{kmj}, s^2_{ktmj} + \sigma^2_{kmj}\right), k = O, A,\ 8 < m \leq 13 \tag{13}$$

$$y_{Qtmj}|\mu_{tj} \overset{\text{ind}}{\sim} \mathsf{N}\left(\mu_{tj} + b_{Qmj}, s^2_{Qtmj} + \sigma^2_{Qmj}\right), m = 13 \tag{14}$$

$$y_{Gtmj}|\mu_{tj} \overset{\text{ind}}{\sim} \mathsf{N}\left(\mu_{tj} + b_{Gmj}, \sigma^2_{Gtmj}\right), m = 10, 11, 12, 13 \tag{15}$$

$$y_{Mj}|\mu_{tj} \overset{\text{ind}}{\sim} \mathsf{N}\left(\mu_{tj}, \sigma^2_{Mj}\right). \tag{16}$$

Prior distributions on the data and process model parameters of each state are specified as before. The full conditional distribution of yield in the j^{th} state, μ_{tj}, resembles

Equation 9. Assuming independence, the collection of state-level crop yields follows a multivariate normal distribution.

$$[\boldsymbol{\mu}_{t\cdot}|\boldsymbol{y}, \boldsymbol{\Theta}_d, \boldsymbol{\Theta}_p] \overset{\text{ind}}{\sim} \mathsf{MVN}\left(\text{vec}\left(\frac{\Delta_{2tj}}{\Delta_{1tj}}\right),\ \text{diag}\left(\frac{1}{\Delta_{1tj}}\right)\right), \tag{17}$$

where $\boldsymbol{\mu}_{t\cdot}$ is the vector of state-level yield parameters. While yield parameters for the region μ_t and states μ_{tj} must respect the balance identity $\mu_t = \sum_j w_{tj}\mu_{tj}$, estimates of parameters $\hat{\mu}_{tj}$ derived under Equation 17 may not, where w_{tj} is weight for state j proportional to harvested acreage. 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 17 on the restriction that their weighted sum is equal to forecasted speculative region yield μ_t, the yield vector for the first $J-1$ states will follow a multivariate normal distribution

$$\left(\mu_{t1}, \mu_{t2}, \ldots, \mu_{t(J-1)}\right) \sim \mathsf{MVN}(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\Sigma}}). \tag{18}$$

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

$$\mu_{tJ} = \mu_t - \frac{1}{w_{tJ}} \sum_{j=1}^{J-1} w_{tj}\mu_{tj}, \tag{19}$$

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 9, Equation 18, and Equation 19 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. Covariate selection

4.1 Covariates in the existing model

In the existing upland cotton model in Section 3, the means of the conditional distributions of parameters μ_t and μ_{tj} are specified as linear combinations of four covariates. These covariates are average July precipitation (pcp_7), average July cooling degree days (cdd_7), crop condition ratings (condGE_30) as of Week 30 and a drought severity index (drght_7). Covariate values for the speculative region are estimated as weighted averages of state-level covariates, where the weights are proportional to harvested acreages.

Covariates in the current model are selected mainly based on expert suggestions on the growth and development processes of crops. Our interest is in determining whether more objectively selected sets of covariates would provide more accurate yield forecasts than covariates in the existing model. In an attempt to select an optimal set of covariates for the upland cotton yield model, we begin with a large pool of potential covariates and apply exploratory analyses, dimension reduction and variable selection techniques to identify an optimal set. Initially, simple exploratory analyses are carried out to eliminate redundant columns from the covariate matrix. The variables are then grouped into similar clusters by using a hierarchical clustering algorithm in the SAS procedure VARCLUS and an optimal variable is selected from each cluster. Spike-and-slab priors (Kuo and Mallick, 1998; George and McCullough, 1993) are specified to the cotton model to identify the most relevant of the covariates selected from clustering analysis. Predictive performances of the

selected set of covariates, the covariates in the existing model and several other covariate sets chosen based on expert suggestions and results from exploratory analysis are compared by using relative differences of the August and September yield forecasts from the May final yield.

4.2 Pool of potential covariates and dimension reduction

Data on several potential covariates are obtained from within NASS, from the National Oceanic and Atmospheric Administration (NOAA) and from the University of Nebraska Lincoln (UNL).

Table 1: Pool of potential covariates

Variable	September
cdd	Cooling degree days
hdd	Heating degree days
tmax	Maximum temperature
tmin	Minimum temperature
tmp	Average temperature
pcp	Average precipitation
sp01	1-month Standardized Precipitation Index
sp02	2-month Standardized Precipitation Index
sp03	3-month Standardized Precipitation Index
zndx	Palmer Z index
pdsi	Palmer Drought Severity Index
phdi	Palmer Hydrological Drought Index
pmdi	Modified Palmer Drought Severity Index
drght	Drought (% Extreme + % Exceptional)
exc	Crop condition: Excellent
condGE	Crop condition: Good+ Excellent
condVP	Crop condition: Poor + Very poor
vp	Crop condition: Very poor
ndvgl	Normalized difference vegetation index

Summaries of crop condition ratings and normalized difference vegetative indices (NDVI) are obtained from NASS sources and data on several weather related variables are extracted from the NOAA and UNL websites. The weekly records of variables on crop condition ratings and average monthly summaries of the weather and NDVI related variables shown in Table 1 are considered as potential covariate values. In the remaining material, the monthly or weekly value of a variable is denoted by adding an underscore and the month or week number to the variable name shown in Table 1. For example, cdd_7 represents average cooling degree days for month 7 for a state or the speculative region.

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. Previous research on the Bayesian hierarchical model (Cruze, 2015, 2016; Cruze & Benecha, 2017, Benecha et al., 2018) as well as exploratory analyses show that covariates have more impact on yield forecasts during the first few months of the crop season and that the impacts of covariates on yield forecasts decrease through the season. Based on these considerations and exploratory analyses, the pool of covariates was reduced to include only the May, June and July values

of the variables shown in Table 1. In addition, some of the variables shown in the table are dropped completely based on correlation analysis.

To reduce the dimension of the covariate matrix, we apply a variable clustering approach that uses a binary and divisive algorithm in the SAS software procedure VARCLUS. In this method, covariates are grouped into similar clusters and a representative covariate is selected from each cluster. As a results, a total of eight potential covariates are selected for further consideration: tmp (average temperature), pcp (average precipitation), zndx (Palmer Z index), pmdi (Modified Palmer drought index), exc (Crop condition rating: Excellent), condGE (Crop condition rating: Good+ Excellent), drght (Drought) and ndvgl (Normalized difference vegetation index). Further reduction in the dimension of the resulting covariate matrix is obtained by including spike-and-slab priors into the Bayesian model as discussed in the next section.

4.3 Applying spike-and-slab priors

Monthly or weekly values of the eight variables picked in Section 4.2 are considered for further filtering by including spike-and-slab priors to the model (Kuo and Mallick, 1998; George and McCulloch, 1993). To implement this approach, parameters related to the regression coefficients are modified, but all other parameters and the structure of the cotton model remain the same. In both the state and regional models, we replace the prior for the regression coefficient associated with covariate j by the priors and hyperpriors (Kou and Mallick, 1998) shown in Equation 20.

$$\begin{aligned} \beta_j &\sim \gamma_j \times \mathsf{N}(0, 1/\tau) \\ \gamma_j &\sim \mathsf{Bernoulli}(p) \\ p &\sim \mathsf{Uniform}(0, 1) \\ \tau &\sim \mathsf{Gamma}(0.001, 0.001) \end{aligned} \tag{20}$$

Two variables are selected based on this approach: condGE_30 (July crop condition rating: Good+ Excellent) and ndvgl_7 (July average NDVI). Model-based forecasts are shown in Figures 3 and 4 respectively for the covariate set $\{$condGE_30, ndvgl_7$\}$ and the existing set of covariates with the ASB and the May final yield estimates. Notice that the two covariate sets provide similar estimates during the last few months of the forecasting season and that the differences in the two sets are mostly in the August and the September forecasts. To further compare the forecasting performances of the existing and the selected sets of covariates, sums of absolute relative differences of the August and September model-based forecasts from the May final yield are calculated as shown in Equation 21.

$$\text{abs.rel.dif}_m = \sum_{t=S}^{T} \frac{|\text{YieldForscast}_{tm} - \text{MayYield}_t|}{\text{MayYield}_t} \tag{21}$$

m = Aug., Sep., T = 2018, S= 2001 & S = 2014

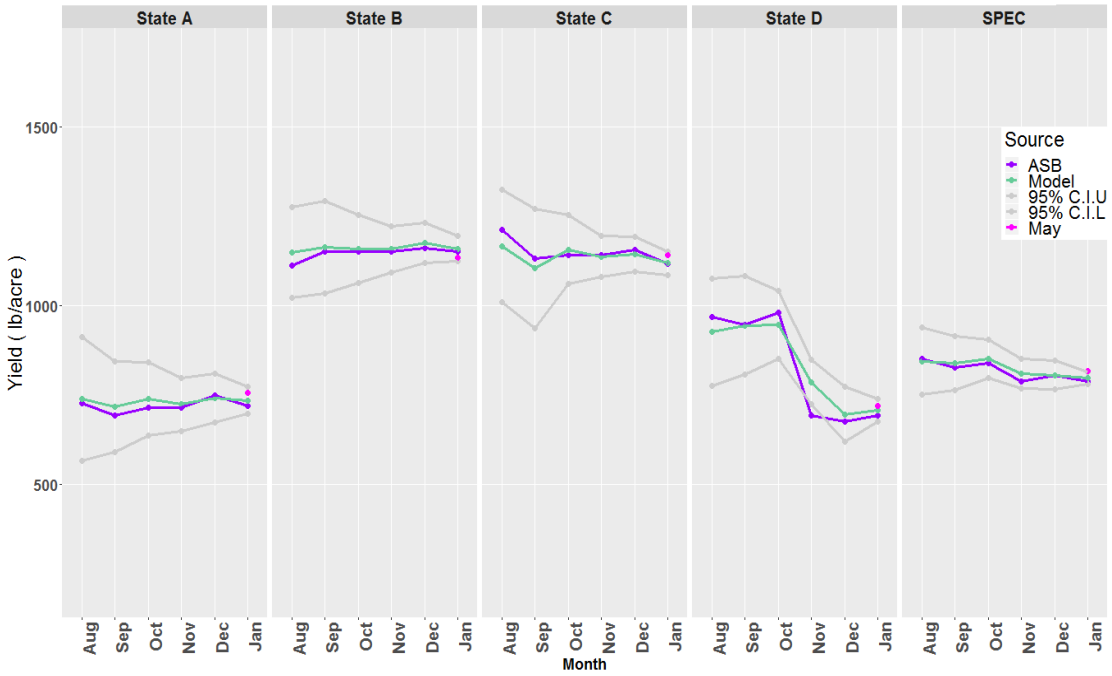

Figure 3: Published and model forecasts based on covariate set {condGE_30,ndvgl_7} for 2018

Figure 4: Published and model forecasts based on covariate set {condGE_30, cdd_7, pcp_7,drght_7 } for 2018

In Table 2, the sums of absolute relative differences of model based forecasts are shown for the speculative region for the years from 2001 to 2018, computed using leave-one-out cross validation (CV), and the the formal model-based yield forecasts for the years from 2014 to 2018. In the latter approach, data from 2001 to the year of interest are used to make forecasts for that year.

Table 2: Sums of absolute relative differences of model estimates

Model	August	September
Leave-one-out CV 2001-2018		
Selected covariates	1.374	1.110
Existing covariates	1.489	1.154
Forecasts for years 2014-2018		
Selected covariates	0.244	0.215
Existing covariates	0.281	0.219

Table 2 shows that the selected covariates provide uniformly smaller absolute relative differences than the existing set of covariates, although the differences are very small.

4.4 Consideration of other potential covariates

On average, the selected set of covariates produce improved forecasts relative to the existing set of covariates. However, it may not necessarily be true that these covariates provide overall optimal forecasts compared to other possible covariate sets. To determine whether there are other sets of covariates that can give more accurate forecasts than the covariate set {condGE_30, ndvgl_7}, a total of 71 covariate combinations are assembled and model-based forecasting is performed for the years from 2001-2018 for each of the covariate sets.

Table 3: Sums of absolute relative differences from 10 top covariate sets

	Sum abs.rel.dif	
Covariates	August	September
{ndvgl_7}	1.368	1.100
{condGE_30, drght_7, ndvgl_7}	1.372	1.098
{condGE_30, ndvgl_7}	1.374	1.110
{condpv_29, ndvgl_7}	1.387	1.112
{zndx_7}	1.395	1.102
{condGE_30, tmp_7, ndvgl_7}	1.396	1.122
{condGE_30}	1.403	1.107
{condGE_30, pcp_7, ndvgl_7}	1.404	1.120
{condGE_30, zndx_7, ndvgl_7}	1.407	1.103
{condpv_29, tmp_7, ndvgl_7}	1.409	1.124

The 71 covariate sets are picked based on expert suggestions and from the results of exploratory data analysis. Table 3 shows that covariate set {ndvgl_7} provides the smallest average relative difference for August and covariate set {condGE_30, drght_7, ndvgl_7} provides the smallest relative difference for September.

Figure 5: Published and model forecasts based on covariate set {condGE_30, drght_7, ndvgl_7} for 2018

Figure 6: Published and model forecasts based on covariate set {condGE_30, cdd_7,pcp_7, drght_7} for 2018

To further compare the top two covariate sets, model-based forecasts are produced based on the two sets for the months from October to January. Overall, the second set of covariates (i.e., {condGE_30, drght_7, ndvgl_7}) provides more accurate forecasts. Figures 5 and 6 show that the covariate set {condGE_30, drght_7 ndvgl_7} and the covariates in existing model (i.e., {condGE_30, cdd_7, pcp_7, drght_7}) provide very close forecasts in October, November, December and January for all states and the speculative region. In

general, the differences between forecasts from the two covariate sets decrease from August to January. The same conclusions can be made about any two covariate sets as the impacts of model covariates on yield decrease though the forecasting season. In particular, covariates have little effects on yield forecasts in December and January mainly because of the availability of data from more surveys that also provide more accurate yield estimates. It can also be seen from Figures 3, 5 and 6 that model-based and ASB yield forecasts in August, September and October are much higher than the forecasts during the remaining months or the May final yield for state D. Such phenomena sometimes happens when a natural disaster hits the crop after data are collected or monthly forecasts are made.

5. Discussion

NASS's Bayesian hierarchical crop yield forecasting models input data from multiple surveys, administrative data and covariates to produce a single forecast for a state or speculative region. The modeled state or regional level yield forecasts can be considered as weighted sums of estimates from the different data sources, including covariates. Model covariates have more impacts on yield forecasts in the first few forecasting months than near the end of the season, mainly because of the availability of data from more surveys that also provide more accurate yield estimates. Covariates in these models are selected mainly based on consideration of the growth and development process of each of the crops.

In this paper, we discussed a more objective approach of selecting covariates for the yield forecasting models with a focus on the upland cotton yield model. We began with a large number of monthly and weekly potential covariates and applied exploratory analyses and a variable clustering method to reduce the dimension of the covariate matrix. Spike-and-slab priors were then included in the model to identify optimal predictors among covariates in the reduced pool. Performances of the selected covariates and those in the existing model are compared for the speculative region using absolute differences of the August and September forecasts relative to the May final yield. As the selected covariates may not necessarily guarantee optimal yield forecasts, we considered several additional covariate sets and compared their forecasting performances with the selected covariates as well as covariates in the existing model. After this step, we selected three covariates for the upland cotton model that provide small relative differences for August and September and perform well for the months from October to January. Overall, our analyses showed that two different covariate sets provide similar forecasts towards the end of the season, and that much of the differences among covariate sets are observed in the first few months of the season. Sometimes early season model based forecasts can be much higher than the May final yield because of natural disasters that impact crop yield and production after forecasts are made.

Acknowledgements

The authors would like to thank Noemi Guindin of USDA NASS Research and Development Division for her input and for creating the map for the speculative region.

Disclaimer

The findings and conclusions in this presentation are those of the authors and should not be construed to represent any official USDA, or U.S. Government determination or policy.

References

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

Benecha, H., Cruze, N., Guindin, N., and Sedransk, N. (2018). Model-Based Crop Yield Forecasting: Adjustment forWithin-State Heterogeneity, Covariate Selection and Variance Estimation. In JSM Proceedings, Survey Research Methods Section. Vancouver, BC: American Statistical Association.

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.

George, I., E. and McCulloch, E., R. (1993). Variable selection via gibbs sampling. *Journal of the American Statistical Association*, 88:881–889.

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

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.

Kuo, L. and Mallick, B. (1998). Variable selection for regression models. *Bayesian Analysis*, 60:65–81.

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.

National Drought Mitigation Center, University of Nebraska-Lincoln (2019). United States Drought Monitor. `https://droughtmonitor.unl.edu/`. Accessed: June 2019.

NOAA National Climatic Data Center (2019). U.S. Climate Divisions: nClimDiv Data Set. `http://www.ncdc.noaa.gov/monitoring-references/maps/us-climate-divisions.php`. Accessed: June 2019.

SAS Institute Inc. 2015. SAS/STAT 14.1 Users Guide. Cary, NC: SAS Institute Inc. (2015). The VARCLUS Procedure. `https://support.sas.com/documentation/onlinedoc/stat/141/varclus.pdf`. Accessed: June 2019.

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.