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

Habtamu Benecha[1], Luca Sartore [1,2]
Nathan B. Cruze[1]

[1]United States Department of Agriculture
National Agricultural Statistics Service (NASS)

[2]National Institute of Statistical Sciences (NISS)

Joint Statistical Meetings
July 29, 2019

Disclaimer

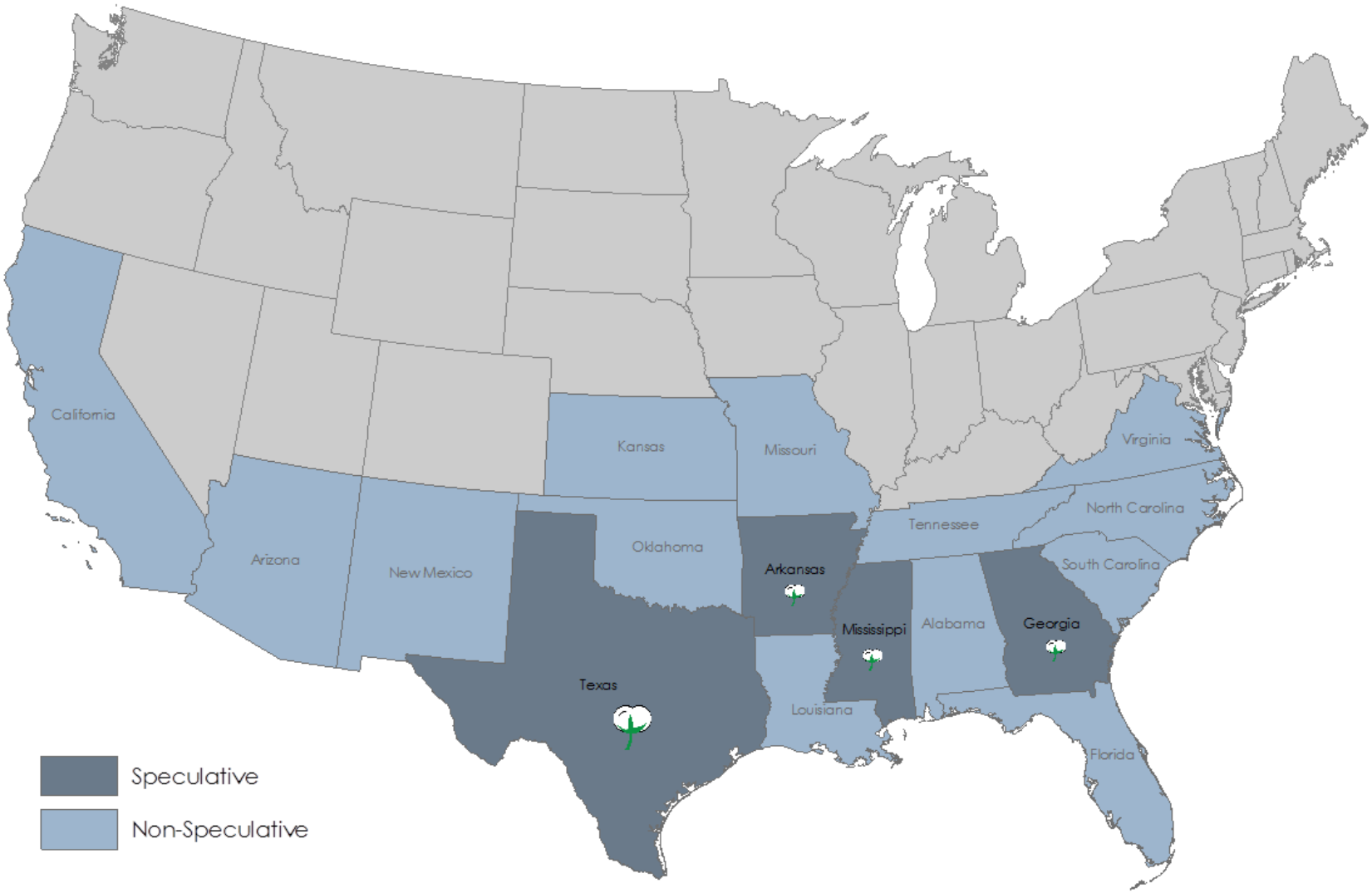
The findings and conclusions in this presentation are those of the authors and should not be construed to represent any official USDA, or U.S. Government determination or policy.

Background and goals

- NASS produces monthly crop yield forecasts
- Official forecasts are consensus estimates of the Agricultural Statistics Board (ASB)
- Recent research in support of the forecasting program
- Bayesian hierarchical models
- Combine data from multiple surveys and covariates

Goal: Which observable covariates are most relevant? Special focus on monthly upland cotton yield forecasting.

Speculative region for upland cotton (2008-2018)



Sources of data

- Objective Yield Survey (**OYS**): monthly field measurements at sampled plots (Sep.-Jan.; Aug.-Jan. prior to 2019)
  - Speculative region
- Agricultural Yield Survey (**AYS**): interview conducted monthly
- December Crops Acreage, Production, and Stocks Survey (**APS**): interview conducted post-harvest, large sample sizes
- Cotton Ginnings (**CG**): projected production data from cotton ginning operations

Bayesian hierarchical model for the Speculative region

(Benecha & Cruze (2018), Adrian (2012), Wang et.al (2012))

Notation

- μ_t–true yield
- y_{ktm}–observed yield
- $k \in \{O, A, Q, G\}$–survey index
- $t \in \{1, ..., T\}$–year index
- $m \in \{8, 9, 10, 11, 12, 13\}$

Stage 1

$$y_{ktm}|\mu_t \overset{ind}{\sim} N\left(\mu_t + b_{km}, s^2_{ktm} + \sigma^2_{km}\right), \quad (1)$$
$$k = O, A, Q, G;\ m = 8, 9, 10, 11, 12, 13$$

$$y_{may}|\mu_t \overset{ind}{\sim} N\left(\mu_t, \sigma^2_{may}\right) \quad (2)$$

Stage 2

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

Existing model Covariates

Covariates for the j^{th} state

$$\mu_{tj} \sim N\left(\boldsymbol{x}'_{tj}\boldsymbol{\beta}_j, \sigma^2_\eta\right)$$

Current model for upland cotton includes:

- cdd_7: Average July cooling degree days (NOAA)
- pcp_7: Average July precipitation (NOAA)
- condGE_30: Crop condition rating % rated excellent + good (NASS); Week 30
- drght_7: July drought severity index (University of Nebraska, Lincoln)

- For the Speculative Region: covariate values are defined as weighted averages of state-level covariate values
- Covariates selection - based on exploratory analysis and knowledge about the growth/development processes of the crop

Pool of potential covariates

Weather, normalized difference vegetation index (NDVI) and crop condition ratings data available

Variable	Description
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)
ave_drght	Another definition of drght
exc	Crop condition: Excellent
condge	Crop condition: Good+ Excellent
condpv	Crop condition: Poor + Very poor
vp	Crop condition: Very poor
ndvgl	Normalized difference vegetation index
ndvmmx	An alternative NDVI variable

Covariate selection: Exploratory analysis, dimension reduction

- Initial exploratory analysis
- Reduce the number of variables by clustering similar columns together
  - Hierarchical clustering
  - Binary and divisive algorithm (SAS/STAT User's Guide 14.1, pages 9787-9817)
- Choose the best variable from each cluster
- Knowledge about the crop to include/exclude variables from the final list
- tmp(average temperature), pcp(average precipitation), zndx(Palmer Z index), pmdi(Modified Palmer drought index), exc(Crop condition: Excellent), condGE(Crop condition: Good+ Excellent), drght(Drought) and ndvgl(Normalized difference vegetation index)

Covariate selection: Spike-and-slab priors

Kou & Mallick (1998); George & McCulloch (1993)

- Insert spike-and-slab priors in the Bayesian model

Corresponding to covariate j, specify

$$\begin{aligned} \beta_j &\sim \gamma_j \times \mathit{Normal}(0, \tau) \\ \gamma_j &\sim \mathit{Bernoulli}(p) \\ p &\sim \mathit{Uniform}(0, 1) \\ \tau &\sim \mathit{Gamma}(0.001, 0.001) \end{aligned} \tag{4}$$

- Simulation studies
- Covariates selected {condGE_30, ndvgI_7 }

Comparison of predictive performances

- Speculative Region: Sums of absolute relative differences of model estimates from May yield ($abs.rel.dif_m$)

$$abs.rel.dif_m = \sum_{t=S}^{T} \frac{|YieldForscast_{tm} - MayYield_t|}{MayYield_t}$$

$$\text{m} = \text{Aug., Sep., Oct., Nov., Dec., Jan,}$$

$$T = 2018, S = 2001 \ \& \ S = 2014$$

Sums of absolute relative differences of model estimates from May yield

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

Covariate selection: considering additional covariate sets

- Best set of covariates from variable selection may not be best for forecasting
  - Consider several additional sets of covariates
- A total of 71 sets of covariates

1. { condpv_29, pcp_7, ndvgl_7 }
2. { condGE_30, pcp_7, zndx_7 }
3. { tmp_7, zndx_7, condGE_30, ndvgl_7, pcp_7 }
4. { condpv_29, zndx_7 }
5. { tmp_7, zndx_7, condpv_29, ndvgl_7 }

.

.

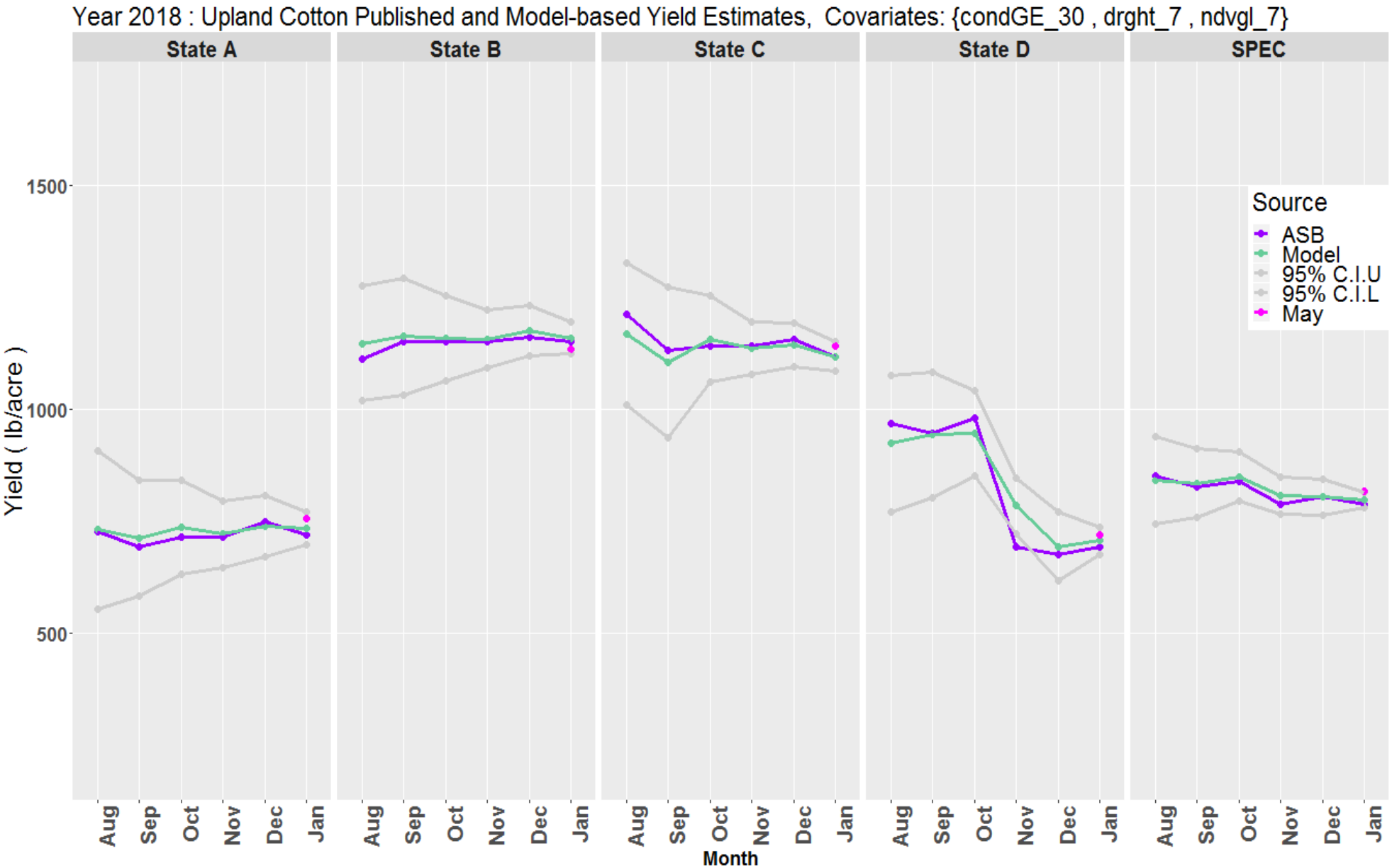
.

71. { tmp_7, zndx_7, condpv_29, ndvgl_7, pcp_7 }

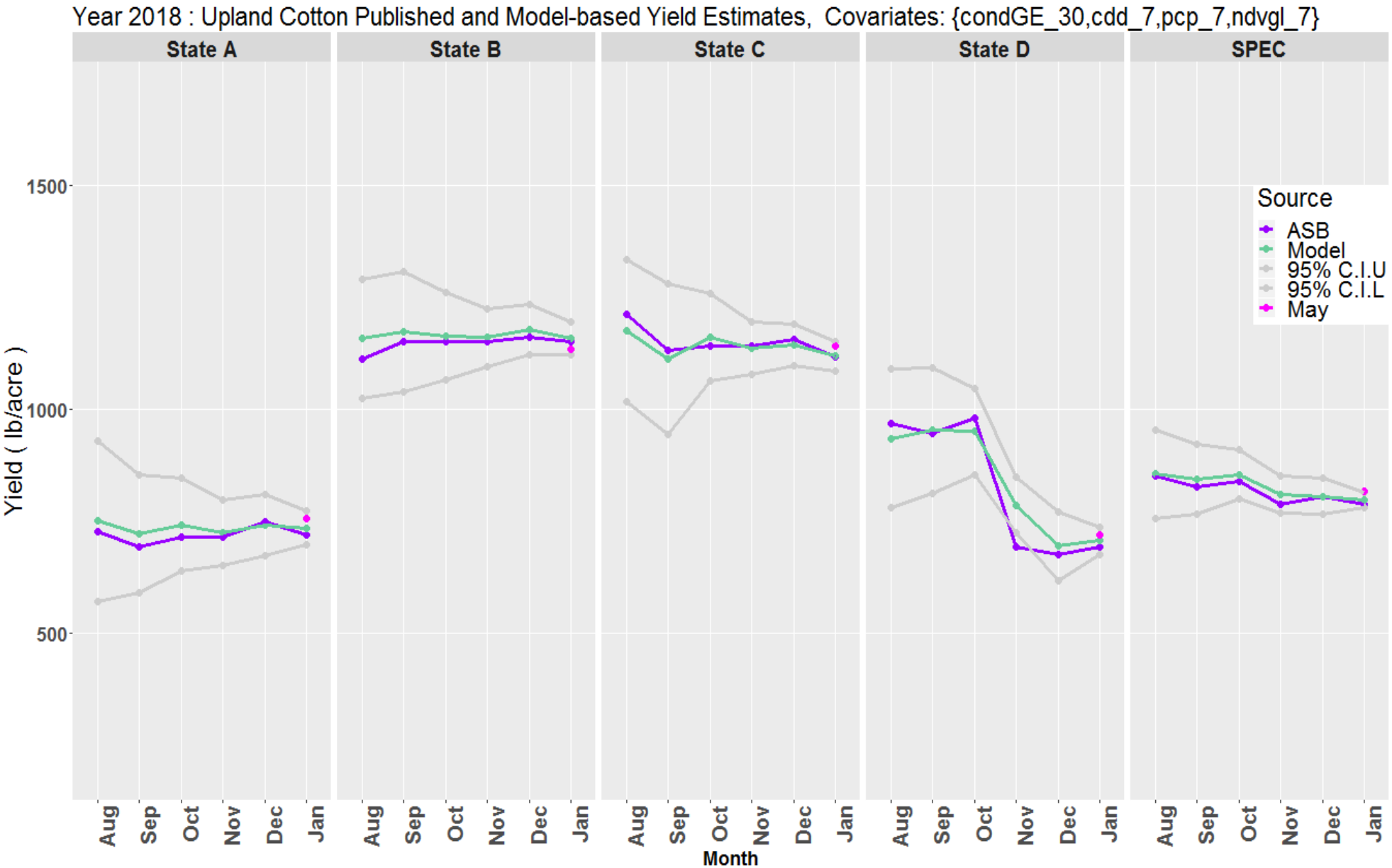
Comparison of predictive performances

Covariates in model	Sum abs.rel.diff	
	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, ave_drght_7, ndvgl_7	1.404	1.120
condGE_30, zndx_7, ndvgl_7	1.407	1.103

Forecasts from {condGE_30,drght_7,ndvgl_7}



Forecasts from existing model: {condGE_30,cdd_7,pcp_7, drght_7 }



Discussion

- Exploratory analysis, cluster analysis, spike-and-slab priors
- NASS crop condition ratings and NDVI are important predictors
- The influence of covariates on yield forecasts decreases from August to January
- Covariates have little impact during the last forecasting months
- August/September forecasts may sometimes be much higher than the final May yield

Select References

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

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

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*, 33:730773.

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

McMaster, G. S. and Wilhelm, W. (1997). Growing degree-days: one equation, two interpretations. Agricultural and Forest Meteorology.

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.

Thank You!

Contact:

habtamu.benecha@usda.gov