Investigating Covariate Selection for a Bayesian Crop Yield Forecasting Model

Habtamu Benecha
habtamu.benecha@nass.usda.gov

Nathan B. Cruze
nathan.cruze@nass.usda.gov

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

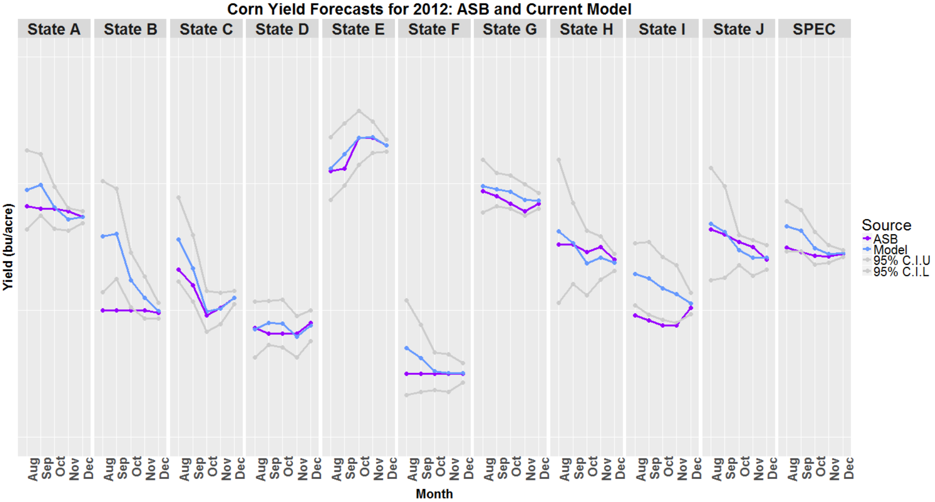
Federal Committee on Statistical Methodology
Research and Policy Conference
March 9, 2018

Background and Research Questions

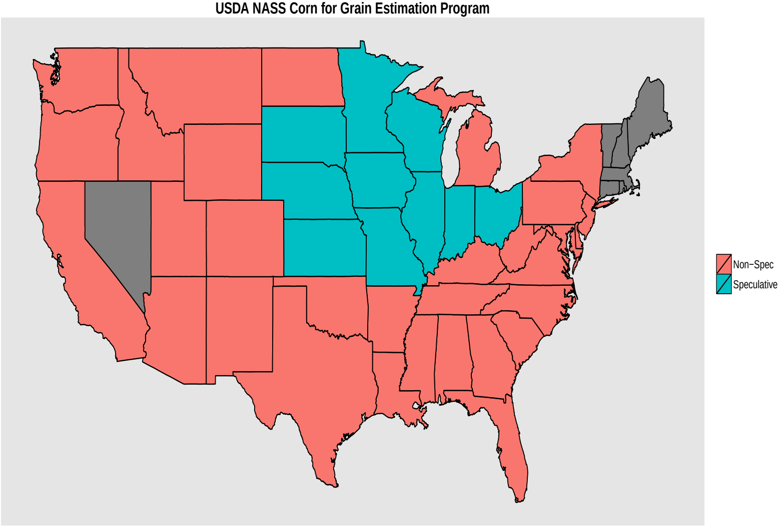
- By mandate, NASS produces monthly crop yield forecasts
- Official forecasts are consensus estimates of the Agricultural Statistics Board (ASB)
- Recent research in support of the forecasting program
- Bayesian hierarchical models
- Combine data from multiple surveys and covariates

Goal: Which observable covariates are most relevant?

Motivating Example: Forecasting in a Drought Year, 2012 Corn



Speculative Region for Corn



NASS Survey Data and Reporting Timeline

Objective Yield Survey (**OYS**)

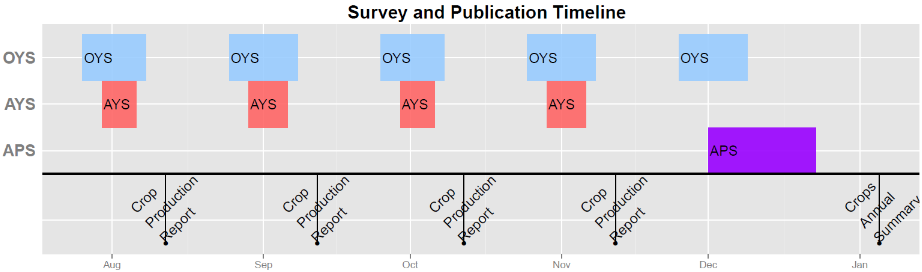
- Field measurements at sampled plots (Aug.- Dec.)

Agricultural Yield Survey (**AYS**)

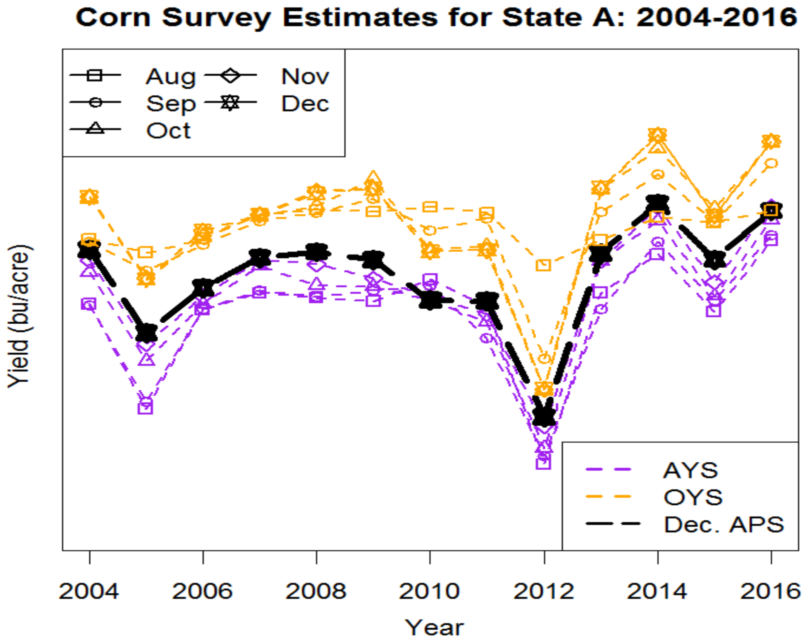
- Interview conducted monthly (Aug.-Nov.)

December Crops Acreage, Production, and Stocks Survey (**APS**)

- Interview conducted post-harvest



Survey Estimates for 2004-2016



Bayesian Hierarchical Model for Speculative Region

Notation

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

Stage 1

$$\begin{aligned} y_{ktm}|\mu_t &\sim \textit{indep } N\left(\mu_t + b_{km}, s_{ktm}^2 + \sigma_{km}^2\right), \\ k &= O, A;\; m = 8, 9, 10, 11, 12 \end{aligned} \tag{1}$$

$$y_{Qt}|\mu_t \sim \textit{indep } N\left(\mu_t, s_{Qt}^2\right) \tag{2}$$

Stage 2

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

Bayesian Hierarchical Model for Speculative Region

Diffuse prior distributions on data and process model parameters

- $\boldsymbol{\Theta}_d \equiv \left(b_{km}, \sigma^2_{km}\right)$
- $\boldsymbol{\Theta}_p \equiv \left(\boldsymbol{\beta}, \sigma^2_{\eta}\right)$

Likelihood function–assuming conditional independence

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

Posterior distribution

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

Bayesian Hierarchical Model–State Level Yield

State-level counterparts indexed by $j \in \{1, 2, ..., J\}$

Unconstrained State Model–Define $\boldsymbol{\mu}_{t\cdot} \equiv (\mu_{t1}, \mu_{t2}, \ldots, \mu_{tJ})$,

$$\boldsymbol{\mu}_{t\cdot} | \boldsymbol{y}, \boldsymbol{\Theta}_d, \boldsymbol{\Theta}_p \sim indep\ MVN\left(vec\left(\frac{\Delta_{2j}}{\Delta_{1j}}\right), diag\left(\frac{1}{\Delta_{1j}}\right)\right) \quad (6)$$

Constrained State Model–Enforce constraint by conditioning state vector $\boldsymbol{\mu}_{t\cdot}$ on $\mu_t = \sum_j w_j \mu_{tj}$

$$\left(\mu_{t1}, \mu_{t2}, \ldots, \mu_{t(J-1)}\right) \sim MVN(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\Sigma}}) \quad (7)$$

$$\mu_{tJ} = \mu_t - \frac{1}{w_{tJ}} \sum_{j=1}^{J-1} w_{tj} \mu_{tj} \quad (8)$$

Covariates for the j^{th} State

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

Current model for corn includes covariates:

- T: Trend
- P: Average July precipitation (NOAA)
- M: Average July temperature (NOAA)
- C: Crop condition rating, % rated excellent + good, Week 30 (NASS)

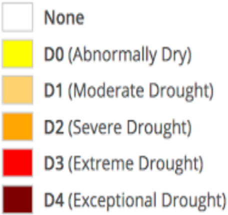
For the Speculative Region: covariate values are defined as weighted averages of state-level covariate values

Additional Covariate

- Early season model-forecasts
- Drought severity index

[1]

  - $D = \%D3 + \%D4$



- Pool of available covariates: $\{T, P, M, C, D\}$
- Potential interactions
- Optimal set of covariates, parsimony

[1] http://droughtmonitor.unl.edu

Challenges in Selecting Appropriate Covariates

- Repeated measure of yield over five months
- Defining a pool of potential covariates
  - Crop-specific knowledge
  - Standard variable selection methods often point to different sets of covariates
    - Step-wise regression
    - LASSO
    - Spike-and-slab regression (Ishwaran and Rao, 2005; Kou and Mallick, 1998), etc
  - Example: {P,M}, {P,M,C,D}, {P,M,D} and {T,P,M,C,D} - 'best' for the Spec-region in 2016
- 'Best' sets of covariates depend on state, year and month

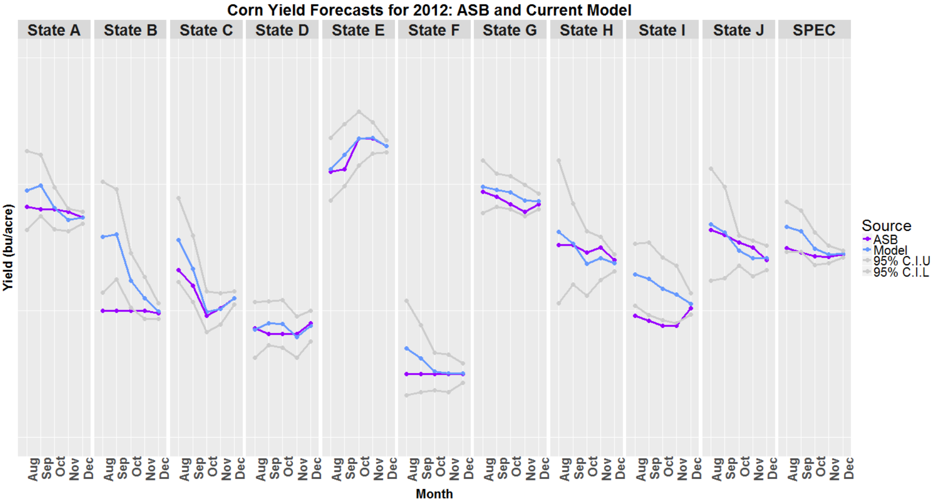
Proposed Approach

1. Start with alternative sets of covariates that are selected most frequently by traditional variable selection methods
2. Fit models for months from August to December and for years 2012, 2013, 2014, 2015 and 2016.
3. Criteria for decision: percent relative difference from Dec. estimates

$$J = \frac{\text{(Aug. forecast - Dec. estimate)}}{\text{Dec. estimate}} \times 100$$

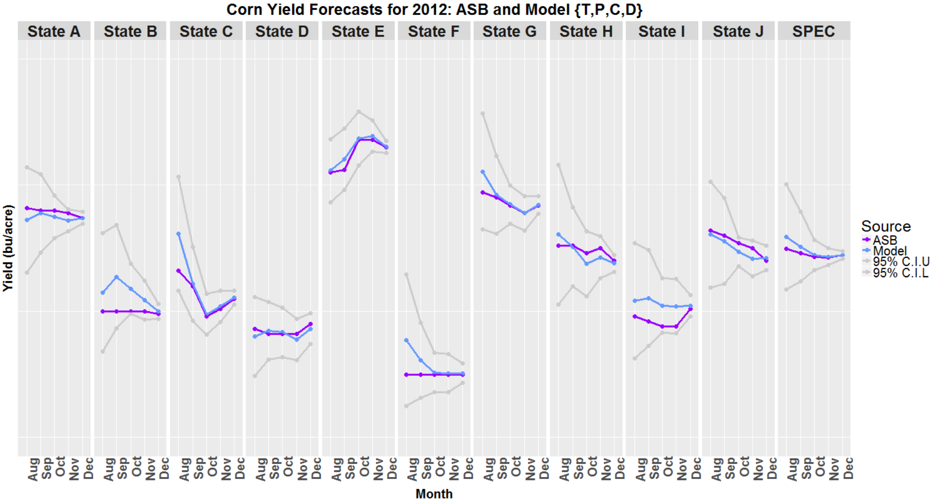
Model Comparison

- A total of 17 covariate combinations
- Subsets of $\{T, P, M, C, D, TD\}$
- Comparisons of models



Model Comparison

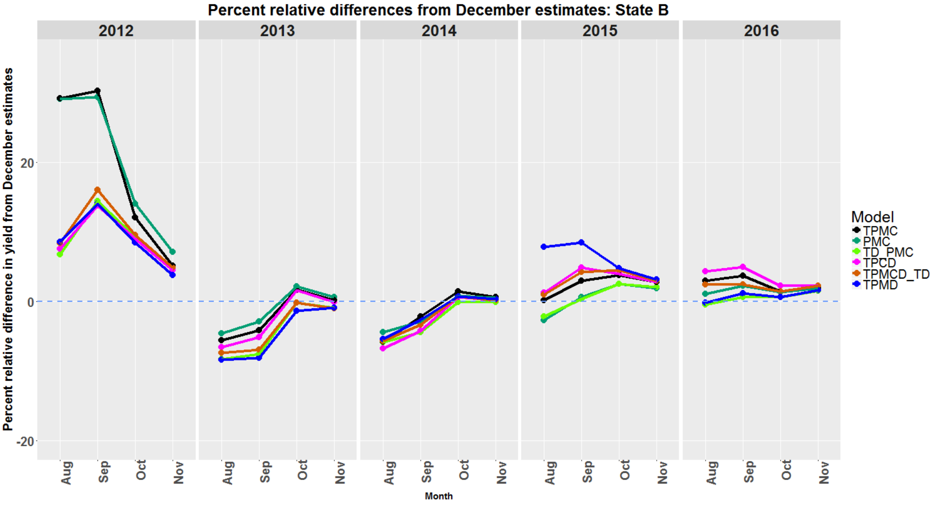
- A total of 17 covariate combinations
- Subsets of $\{T, P, M, C, D, TD\}$
- Comparisons of models



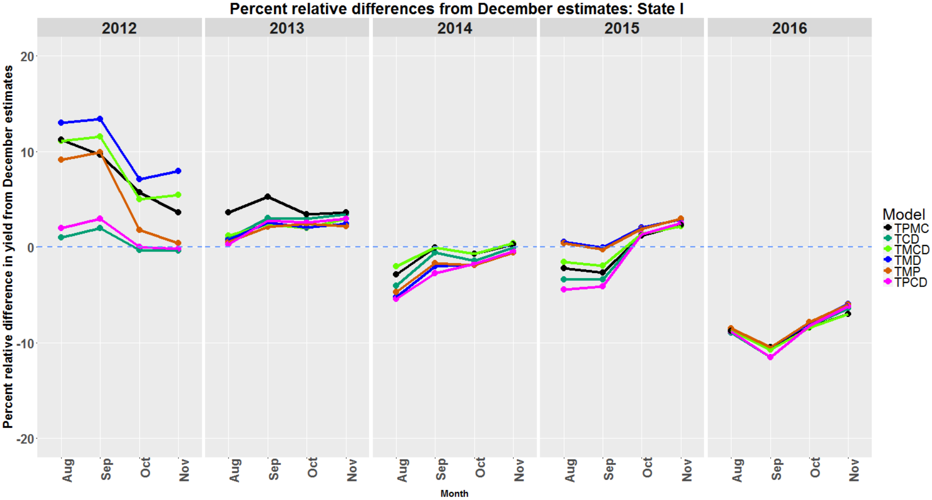
Variables with smallest percent relative differences

	Covariate-sets		
State	Without Drought	With Drought as Main Effect	With Interaction (D*T)
A	-	'13, '15	'12, '14, '16
B	'13, '14, '15	'16	'12
C	'13	'12, '15, '16	'14
D	'12	'13, '14, '15	'16
E	'14	'13, '15, '16	'12
F	'12, '14	'13, '15	'16
G	'12, '13	'14, '15, '16	-
H	'14	'13	'12, '15, '16
I	'12, '15	'13, '14, '16	-
J	'15	'13	'12, '14, '16

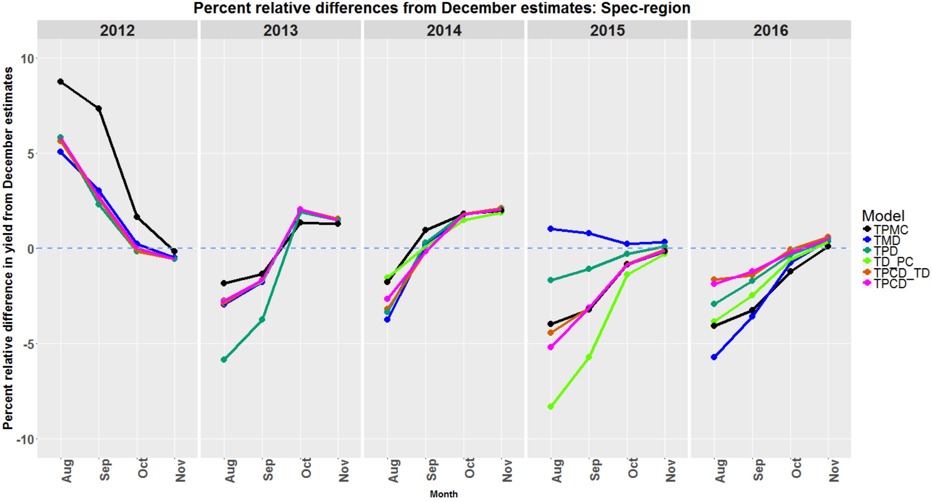
Model Comparison for State B



Model Comparison for State I



Model Comparison for the Spec-region



- DIC values for December 2016 corresponding to covariate-sets {T,P,M,C}, {T,M,D} , {T,P,D} , {P,C, T*D} ,{T,P,C,D,T*D} ,{T,P,C,D} are 162.92, 163.06, 163.07, 162.93, 163.15 and 163.02 respectively.

Conclusions

Investigated sensitivity of model forecasts to linear model specification

- Inclusion of a 'drought' covariate improved early yield forecasts
- No one-size fits all set of covariates
- State-specific covariates may be considered

Contact:
habtamu.benecha@nass.usda.gov

Select 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.

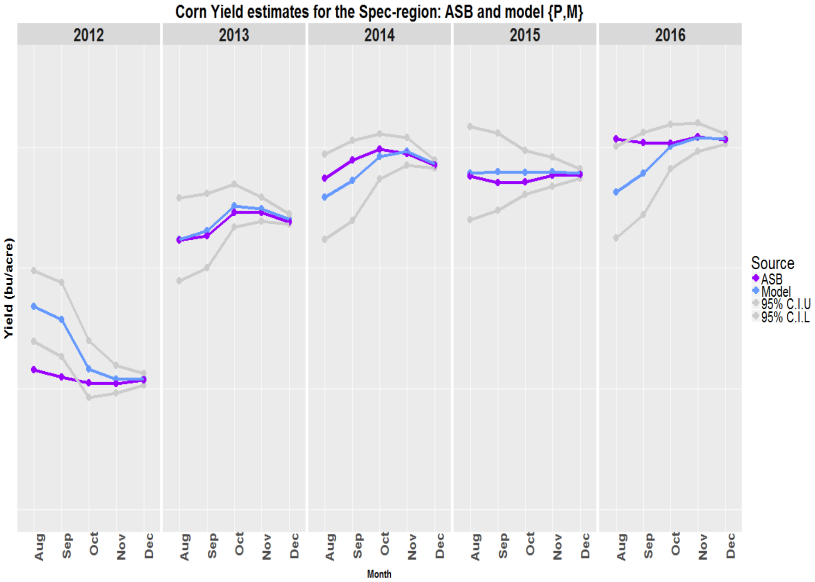
Cruze, N. B. and Benecha, H. (2017). A Model-Based Approach to Crop Yield Forecasting. In JSM Proceedings, Section on Bayesian Statistical Science. Alexandria, VA: American Statistical Association.

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.

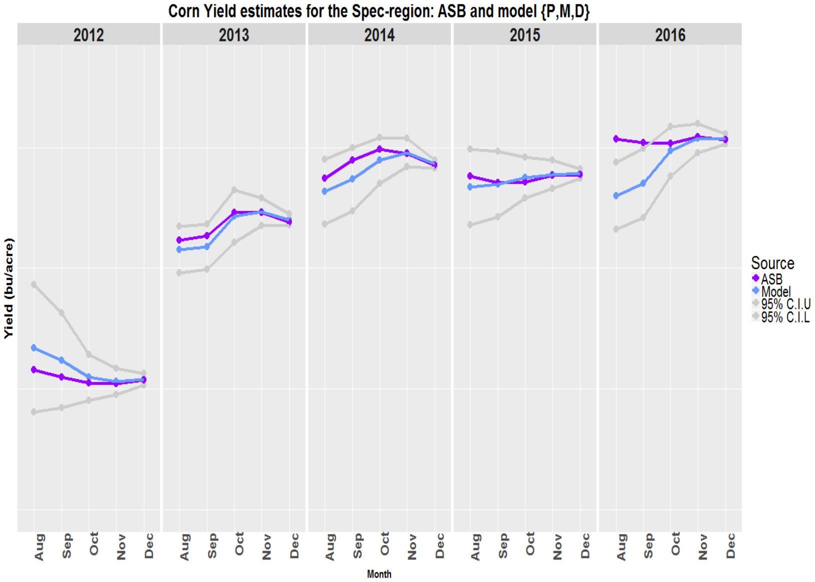
Nandram, B. and Sayit, H. (2011). A Bayesian analysis of small area probabilities under a constraint. *Survey Methodology*, 37:137–152.

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.

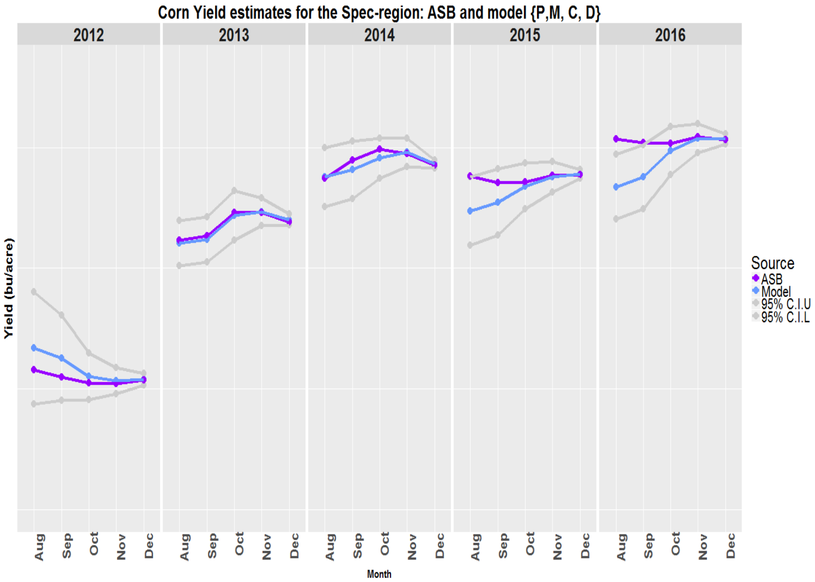
Supplementary material 1



Supplementary material 2



Supplementary material 3



Supplementary material 4