Comparison of Mixed Modeling Approaches to County-Level Crop Area Estimation Using Multiple Data Sources

Michael E. Bellow

USDA/NASS, 1400 Independence Ave., SW, Rm. 6407B
Washington, DC 20250

Efficient estimation of crop parameters at the county (small domain) level is an important priority for the USDA's National Agricultural Statistics Service (NASS). This paper focuses on three mixed modeling approaches to county-level estimation of crop planted or harvested area where survey reported values are fit to unit (farm) and area (county) level covariates: 1) an empirical best linear unbiased predictor (EBLUP) model, 2) an adaptive empirical best prediction (AEBP) model, and 3) a log-transformed empirical best prediction model. In a simulation study involving corn and soybean planted area in Ohio and South Dakota for 2018, the three estimators are compared using data from NASS's County Agricultural Production Survey (CAPS) and auxiliary data sources. Control data from NASS's list sampling frame are used as the unit-level covariate while two options are considered for the area-level covariate: 1) Farm Service Agency (FSA) planted acreage, and 2) satellite-based pixel counts obtained from NASS's Cropland Data Layer. Since the unit-level covariate is missing for a subset of the list frame records, regression synthetic estimation is applied in that portion of the population to ensure complete coverage.

1. Introduction

The National Agricultural Statistics Service (NASS) produces small area statistics on some crop and livestock commodities. A small area refers to a geographical region (e.g., a U.S. county) for which limited information is available from the primary source of data. County-level agricultural statistics are used for a number of applications including regional planning and fund allocation in government programs, setting of crop insurance premiums and agronomics research. The importance of producing accurate, defensible county-level estimates has motivated NASS to maintain an active research program in this area (Cruze, et al., 2019).

NASS's main source of data for commodity estimation has always been surveys of farmers, ranchers and agribusiness managers who provide requested information on a voluntary, confidential basis. In general, traditional direct methods that utilize only small area specific data from surveys designed for higher levels of aggregation (e.g., states) have been unreliable due mainly to small sample sizes in the areas of interest. In addition, effects of different nonsampling errors such as coverage and nonresponse can be severe. Even after combining data from multiple surveys, direct methods have often fallen short of producing adequate small area statistics. To improve on direct estimators, several indirect and model-based methods have been proposed. These estimation procedures use implicit or explicit models that borrow strength from related resources such as administrative records, previous year survey estimates, NASS list sampling frame control data, agricultural census data and earth observing satellite data.

Prior to the implementation of CAPS, county estimation used data from sample surveys designed for estimating values at higher levels of aggregation in conjunction with ancillary data sources. Although CAPS is designed specifically for county-level estimation, there is still potential for improvement via the application of model-based small area estimation methodology.

Previously, an empirical best linear unbiased predictor (EBLUP) estimator based on the Battese-Harter-Fuller (BHF) model (Battese et al., 1988) and involving two covariates (one at the unit-level and the other at the area-level) was evaluated for harvested acreage (Bellow and Lahiri, 2011). In an empirical study conducted over seven states in the Midwestern grain belt region of the U.S., the BHF estimator was found to be more accurate overall for corn and soybeans than five competing estimators when NASS official county-level estimates were used as the gold standard for comparison. A subsequent three-state (Illinois, Maryland, Tennessee) study showed a newly proposed mixed-model estimator known as *adaptive empirical best prediction* (AEBP) to be superior (in general) to both *BHF* and a log-transformed estimator called LEBP (Bellow and Lahiri, 2012). The term 'adaptive' refers to the fact that a hyperparameter is estimated by the model fitting algorithm as opposed to being preset to a fixed value. Neither of these two empirical studies involved the use of simulation.

In this paper, we use Monte Carlo simulation to evaluate the two estimators just referred to (BHF and AEBP) and a third which is a special case of the model-based estimator developed by Berg and Chandra (2014). The latter estimator (referred to as BC*)* is based on a log-transformed model related to LEBP.

In our application, the areas are counties and the population units are farming operations. The unit (farm) level covariate employed is derived from the *size variable*, a measure of planted area for the crop of interest over recent survey years based on NASS list frame control data. A missing value for this variable could arise if the farm in question is in the list frame sample for the current year but not the previous year, while a zero value could occur if the crop was never planted on the farm over recent years. Two available choices for the source of the area (county) level covariate are:

1) Farm Service Agency (FSA) planted acreage – estimates of planted acreage derived from reports submitted by farm operators to their local FSA offices, and

2) Remote Sensing (RS) pixel counts – counts of satellite pixels classified to crops in NASS's Cropland Data Layer (CDL).

The FSA planted acreage figures are normally available to NASS in time for use in operational county-level estimation (which involves a board process). NASS's Cropland Data Layer (CDL) is a georeferenced, crop-specific data layer of land cover (including crop types) created annually for the continental U.S. using a combination of Landsat 8, Deimos-1, UK-DMC2 and Sentinel-2 satellite imagery and agricultural ground reference data (Boryan et al., 2011).

Although a full scale comparison between the impacts of using the FSA and RS covariates is beyond the scope of this paper, preliminary evidence suggests that the choice of one or the other makes little difference in terms of the resulting model-based estimates. For that reason, the FSA covariate is used exclusively in our estimator comparison study.

Due to incomplete availability of the unit-level covariate over the entire population, estimation of total planted acreage must be of the *hybrid* variety. The population is subdivided into two poststrata: A (unit-level covariate available) and B (unit-level covariate missing). Mixed effects estimation is applied in poststratum A and an area-level estimator in poststratum B.

In Section 2, we describe the three mixed effects models for county-level acreage estimation and the associated estimation procedures in detail. Section 3 discusses a two-state (Ohio and South Dakota) simulation study comparing three hybrid estimators for corn and soybeans in 2018 (one each corresponding to BHF, AEBP and BC in poststratum A and regression synthetic estimation in poststratum B). Section 4 provides a summary and discussion of potential avenues for future research.

2. Mixed Effects Models for County-Level Crop Area Estimation

In this section, we first introduce a general mixed effects model and then discuss the three specific versions noted above and their associated estimation procedures.

Linear Mixed Effects Model

The general mixed effects model with *m* covariates is expressed as:

$$f(y_{ij}) = \boldsymbol{z}_{ij}'\boldsymbol{\beta} + v_i + e_{ij} \quad (i = 1, \ldots, L;\ j = 1, \ldots, N_i)$$

where:

f (t) = t or log(t),

y_{ij} = value of dependent variable for area *i*, population unit *j,*

$\boldsymbol{z}_{ij}' = (1,\ z_{1ij}, \ldots, z_{mij})$ - vector of covariates for area *i*, population unit *j*,

$\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_m)'$ - vector of model regression parameters,

L = number of areas to be estimated,

N_i = number of population units in area *i,*

v_i = effect for area *i,*

e_{ij} = random error for area *i*, population unit *j,*

$(v_i, e_{ij}) \sim N(0, \mathrm{diag}(\sigma_v^2, \sigma_{eij}^2))$.

In our specific application involving a single unit-level and area-level covariate and the two poststrata defined in Section 1, the model reduces to:

$$f(y_{ij}) = \beta_0 + \beta_1 z_{1ij} + \beta_2 z_{2i} + v_i + e_{ij}$$

where:

y_{ij} = true area (planted or harvested) of crop in county *i*, population unit *j*.

Let:

x_{1ij} = value of list frame size variable in county *i*, population unit *j*,

N_i = number of population units in county *i,*

n_{Ai} = number of sample units in poststratum *A*, county *i,*

$x_{2i}^{(FSA)}$ = (FSA planted acreage in county *i) / N_i,*

$x_{2i}^{(RS)}$ = (RS pixel count in county *i) / N_i.*

The models referred to as Battese-Harter-Fuller (BHF), adaptive empirical best prediction (AEBP) and Berg-Chandra (BC) are defined by specific choices of f(t), the two covariates and the random error variance structure. All three are fit using the *restricted maximum likelihood* (REML) method. Before proceeding further, we augment the notation for regression parameters and county effects so as distinguish among the three models:

$\beta_{Ak}^{(M)}$ = regression parameter associated with model M in poststratum *k* (*k* = 0, 1, 2),

$v_i^{(M)}$ = county effect associated with model M (*i* = 1, …, L).

Battese-Harter-Fuller (BHF) Model

If f(t) = t, the unit-level covariate is x_{1ij} , the area-level covariate is either $x_{2i}^{(FSA)}$ or $x_{2i}^{(RS)}$ and $\sigma_{eij}^2 = \sigma_e^2$ (constant over population units), we obtain a special case of the model first proposed by Battese, Harter and Fuller (1988) for a related agricultural application.

$$y_{ij} = \beta_{A0}^{(BHF)} + \beta_{A1}^{(BHF)} x_{1ij} + \beta_{A2}^{(BHF)} x_{2i} + v_i^{(BHF)} + e_{ij}$$

Adaptive Empirical Best Prediction (AEBP) Model

The specifications of this model are identical to those for BHF except that the random error variance is assumed to be a function of the unit-level covariate (crop size):

$$\sigma_{eij}^2 = x_{1ij}^{\delta} \sigma_e^2 \quad (\delta > 0)$$

AEBP is *adaptive* since the hyperparameter δ is estimated from the data by the REML algorithm which searches for the value that minimizes the *Bayesian Information Criterion*.

Berg-Chandra (BC) Model

If *f(t) = log(t)*, the unit-level covariate is $log(x_{1ij})$, the area-level covariate is either $log(x_{2i}^{(FSA)})$ or $log(x_{2i}^{(RS)})$ and $\sigma_{eij}^2 = \sigma_e^2$, we obtain a version of a model proposed by Berg and Chandra (2014):

$$log(y_{ij}) = \beta_{A0}^{(BC)} + \beta_{A1}^{(BC)} log(x_{1ij}) + \beta_{A2}^{(BC)} log(x_{2i}) + v_i^{(BC)} + e_{ij}$$

Note that this log-transformed model can only be fit using survey records with positive acreage, thus introducing an additional bias that (unless all survey records are positive) must be adjusted for in the estimation process.

The procedure for computing the poststratum *A* component of hybrid county-level (planted or harvested) area estimates using the BHF, AEBP or BC model is as follows:

1) Fit survey acreage values to covariates within sample units in poststratum *A* using the REML algorithm to obtain estimates of model parameters and county effects.

2) Compute estimates of average (per population unit) and total crop area in poststratum *A*, county *i* (respectively) as:

$$\hat{y}_{Ai}^{(M)} = \hat{\beta}_{A0}^{(M)} + \hat{\beta}_{A1}^{(M)} \bar{x}_{1i} + \hat{\beta}_{A2}^{(M)} x_{2i} + \hat{v}_i^{(M)} \qquad \text{(M = BHF or AEBP)}$$

$$\hat{y}_{Ai}^{(BC)} = \rho_i \{ \sum_{s_{A(\sim i)}^+} \exp[\hat{\beta}_{A0}^{(BC)}(1-\hat{\gamma}_i) + \hat{\beta}_{A1}^{(BC)}(x_{1ij} - \hat{\gamma}_i \bar{x}_{1i})) + \hat{\beta}_{A2}^{(BC)} x_{2i}(1-\hat{\gamma}_i) + \hat{\gamma}_i \bar{I}_{Ai}] + n_{Ai}^+ \bar{y}_{Ai} \}$$

$$\hat{Y}_{Ai}^{(M)} = N_{Ai}\, \hat{y}_{Ai}^{(M)} \qquad \text{(M = BHF, AEBP or BC).}$$

where:

$\hat{\beta}_{Ak}^{(M)}$ = estimate of $\beta_{Ak}^{(M)}$ computed by REML algorithm,

$\hat{v}_i^{(M)}$ = estimate of $v_i^{(M)}$ computed by REML algorithm (i = 1, …, L),

ρ_i = adjustment term,

s_A^+ = set of sampled units with positive acreage of crop in poststratum A, county *i,*

$s_{A(\sim i)}^+$ = set of sampled units with positive acreage of crop in poststratum A, counties other than *i,*

n_{Ai}^+ = number of sampled units with positive acreage of crop in poststratum A, county *i,*

$\hat{\beta}_{Ak}^{(BC)}$ = estimate of $\beta_{Ak}^{(BC)}$ computed by fitting *BC* model using REML in poststratum A *(k* = 0, 1, 2*),*

$\hat{\sigma}_v^2$, $\hat{\sigma}_e^2$ = estimates of σ_v^2 and σ_e^2 (computed by *REML* algorithm)

$\hat{\gamma}_i = \hat{\sigma}_e^2/(\hat{\sigma}_e^2 + n_{Ai}^+\hat{\sigma}_v^2)$,

$\bar{x}_{1i}$ = population mean of unit-level covariate (x_{1ij}) in poststratum *A*, county *i,*

N_{Ai} = number of population units in poststratum A, county *i,*

$\bar{l}_{Ai} = (1 / n_{Ai}^+) \sum_{j \in s_{Ai}^+} \log(y_{ij})$ if $n_{Ai}^+ > 0$,
$= 0$ otherwise.

3) Compute *regression synthetic* (*RGS*) estimates of average and total crop acreage in poststratum *B*, county *i* as:

$\hat{y}_{Bi}^{(RGS)} = [\sum_{k=1}^{L} n_{Bk}\, \bar{y}_{Bk} \;/\; \sum_{k=1}^{L} n_{Bk}\; x_{2k}]\, x_{2i}$,

$\hat{Y}_{Bi}^{(RGS)} = N_{Bi}\, \hat{y}_{Bi}^{(RGS)}$

where:

n_{Bi} = number of sample units in poststratum B, county *i,*

$\bar{y}_{Bi}$ = sample mean survey acreage for poststratum B, county *i,*

N_{Bi} = number of population units in poststratum B, county *i.*

4) Compute hybrid estimate of total and average crop area in county *i* as:

$\hat{Y}_i^{(M/RGS)} = \hat{Y}_{Ai}^{(M)} + \hat{Y}_{Bi}^{(RGS)}$,

$\hat{y}_i^{(M/RGS)} = \hat{Y}_i^{(M)} / N_i \quad (i = 1, \ldots, L)$

where:

M = model used in poststratum *A (*BHF, AEBP or BC).

The hybrid estimator that uses model M in poststratum A as is denoted by M/RGS.

3. Simulation Study Comparing Estimators

This section describes a simulation study aimed at comparing efficiency properties of hybrid estimators associated with the three mixed model estimators BHF, AEBP and BC. The study was conducted for corn and soybeans in Ohio and South Dakota for the 2018 crop year.

In order to avoid negative replicates, the Berg-Chandra model was chosen as the simulation model in poststratum A. We should note that since BC/RGS is one of the hybrid estimators being evaluated, there is a possibility of the simulation results being biased in favor of that estimator. The procedure employed to carry out the simulations is as follows:

1) Compute simulation parameter estimates $\hat{\beta}_{Ak}^{(BC)}$ ($k = 0, 1, 2$) and $\hat{\sigma}_{Ae}^2$ by fitting the BC model via REML using sampled units with positive reported acreage in poststratum A.

2) Compute simulation parameter estimates $\hat{\beta}_{Bk}$ ($k = 0, 2$) and $\hat{\sigma}_{Be}^2$ by fitting a simple linear regression model relating survey reported acreages to the area-level covariate in poststratum B.

3) Generate simulated crop acreage values for all population farms in poststratum A using the following formula:

$$y_{Aij}^{(sim)} = \hat{\beta}_{A0}^{(BC)} + \hat{\beta}_{A1}^{(BC)} x_{1ij} + \hat{\beta}_{A2}^{(BC)} x_{2i} + \hat{v}_i^{(BC)} + e_{Aij}^{(sim)}$$

where:

$y_{Aij}^{(sim)}$ = simulated crop acreage in poststratum *A*, county *i*, farm *j*,

$e_{Aij}^{(sim)}$ = value of e_{Aij} simulated from $N(0, \hat{\sigma}_{Ae}^2)$ distribution.

4) Generate simulated crop acreage values for all population farms in poststratum B as:

$$y_{Bij}^{(sim)} = \hat{\beta}_{B0}^{(REG)} + \hat{\beta}_{B2}^{(REG)} x_{2i} + e_{Bij}^{(sim)}$$

where:

$\hat{\beta}_{B0}$, $\hat{\beta}_{B2}$ = ordinary least squares estimates of $\beta_{B0}^{(REG)}$ and $\beta_{B2}^{(REG)}$,

$e_{Bij}^{(sim)}$ = value of e_{Bij} simulated from $N(0, \hat{\sigma}_{Be}^2)$ distribution.

5) Combine the sets of simulated crop acreage values from steps 3 and 4 to generate the overall simulated population from which 'truth' values are derived by summing and averaging statistics over counties:

$$y_{ij}^{(sim)} = y_{Aij}^{(sim)} + y_{Bij}^{(sim)}$$

The next five steps are carried out for each replication *r* ($r = 1, \ldots, N_{rep}$):

6) Select a sample of preset size n_{smp} from the simulated population using simple random sampling.

7) Fit models of interest (BHF, AEBP, and BC) using the poststratum A subset of the sample to derive estimates of the model parameters.

8) Compute poststratum A estimates for each model using the appropriate formulas from Section 2:

$\hat{y}_{(r)Ai}^{(M)}$ = model M estimate of average planted acreage per population unit in

poststratum A.

9) Compute poststratum B estimates using the formula for RGS estimation from Section 2:

$\hat{y}_{(r)Bi}^{(RGS)}$ = regression synthetic estimate of average planted acreage per population unit in poststratum B.

10) Combine poststrata A and B estimates to obtain overall (hybrid) replication *r* county-level estimates of average planted acreage:

$$\hat{y}_{(r)i}^{(H)} = [N_{Ai}\hat{y}_{(k)Ai}^{(M)} + N_{Bi}\hat{y}_{(k)Bi}^{(RGS)}] / N_i$$

11) After all replications have been completed, compute county-level performance metrics for each model based on replicated estimates of planted acreage and 'true' planted acreage values from the simulated population.

The following county-level performance metrics are used in the study (H refers to the specific hybrid estimator, e.g., BHF/RGS*)*:

1) Bias

$$[Bias]_i^{(H)} = \sum_{k=1}^{N_{rep}} (\hat{y}_{(k)i}^{(H)} - \bar{y}_i^{(sim)}) / N_{rep}$$

where:

$$\bar{y}_i^{(sim)} = \sum_{j=1}^{Ni} y_{ij}^{(sim)} / N_i$$

2) Relative Bias

$$[RB]_i^{(H)} = [Bias]_i^{(H)} / \bar{y}_i^{(sim)}$$

3) Absolute Bias

$$[AB]_i^{(H)} = | [Bias]_i |$$

4) Standard Deviation

$$[SD]_i^{(H)} = [\sum_{k=1}^{N_{rep}} (\hat{y}_{(k)i}^{(H)} - \bar{y}_i^{(H)})^2 / (N_{rep} - 1)]^{1/2}$$

where:

$$\bar{y}_i^{(H)} = \sum_{j=1}^{N_{rep}} \hat{y}_{(k)i}^{(H)} / N_{rep}$$

5) Coefficient of Variation

$$[CV]_i^{(H)} = [SD]_i^{(H)} / \bar{y}_i^{(H)}$$

6) Root Mean Squared Error

$$[RMSE]_i^{(H)} = [\sum_{k=1}^{N_{rep}} (\hat{y}_{(k)i}^{(H)} - \bar{y}_i^{(sim)})^2 / N_{rep}]^{1/2}$$

7) Relative Root Mean Squared Error

$$[RRMSE]_i^{(H)} = [RMSE]_i^{(H)} / \bar{y}_i^{(H)}$$

Table 1 provides data on number and percent of population units in poststrata A and B for each of the four state/crop combination considered in the study. Note that the proportion of units in poststratum B (crop size variable missing) ranged from 6 to 19 percent.

Table 1. Number and Percent of Population Units by Poststrata

State	Crop	Poststratum		
		A	B	Combined
Ohio	Corn	23,134 (82%)	5,074 (18%)	28,208
	Soybeans	22,820 (81%)	5,388 (19%)	28,208
South Dakota	Corn	11,612 (94%)	751 (6%)	12,363
	Soybeans	10,490 (82%)	2,283 (18%)	12,773

In order to sufficiently minimize the variation due to simulation, the number of replications used for each case (state/crop/estimator combination) was set to the smallest multiple of 500 greater than or equal to the square of the number of counties to be estimated. The resulting number of replications in the study were 8,000 for both corn and soybeans in Ohio (87 counties), 3,000 for corn in South Dakota (52 counties) and 3,500 for soybeans in South Dakota (57 counties). For each case, the same strings of random number seeds were generated to simulate the random errors for all three estimators (although strings were unrelated across cases). The sample sizes per replication were 1100 for both Ohio cases and 500 for both South Dakota cases (approximately four percent of the total number of population units). The method of sampling in this study is not intended to emulate NASS operational practice where *maximal Brewer selection* (also known as *multivariate probability proportional to size sampling*) is used to select the CAPS samples (Kott and Bailey, 2000). Since the survey reported acreages were positive for all population units with non-missing value of the unit-level covariate (crop size) in all cases, the adjustment term for the BC estimator was always equal to one.

Table 2 shows for each hybrid estimator the seven efficiency metrics defined earlier in this section (averaged over counties) by state and crop. In each row, an asterisk appears next to the 'best' of the three values (closest to zero for bias and relative bias, lowest for the other five metrics). Note that for corn in Ohio, AEBP/RGS and BC/RGS were best for three metrics each and BHF/RGS for one; for soybeans in Ohio, AEBP/RGS was best for four metrics and BC/RGS for the other three. For corn in South Dakota, AEBP/RGS was best for four metrics and BHF/RGS for the other three; for soybeans in South Dakota, AEBP/RGS was best for five metrics with BHF/RGS and BC/RGS best for one metric

each. For corn (both states combined), AEBP/RGS was best for seven state/metric combinations, BHF/RGS for four and BC/RGS for three. For soybeans (both states combined), AEBP/RGS was best for nine state/metric combinations, BC/RGS for four and BHF/RGS for one. Overall, AEBP/RGS was best for 16 state/crop/metric combinations (14 involving variability metrics), BC/RGS for seven and BHF/RGS for five.

Table 2. Performance Metrics for Hybrid Estimators by State and Crop

State	**Crop**	**Metric**	**Hybrid Estimator**		
			BHF/RGS	AEBP/RGS	BC/RGS
Ohio	Corn	Bias	-1.27	9.31	- 0.002*
		RB	-0.017*	0.052	-0.02
		AB	16.7	21.0	15.9*
		SD	60.9	59.5*	60.5
		CV	0.364	0.314*	0.346
		RMSE	65.9	65.5	65.3*
		RRMSE	0.403	0.36*	0.388
	Soybeans	Bias	-5.12	12.81	-4.59*
		RB	-0.126	-0.059*	-0.216
		AB	25.7	36.7	25.7*
		SD	55.0	52.9*	55.3
		CV	0.287	0.209*	0.366
		RMSE	66.3	68.9	66.1*
		RRMSE	0.387	0.337*	0.51
South Dakota	Corn	Bias	1.92*	32.4	13.1
		RB	0.006*	0.064	0.016
		AB	33.4*	46.8	40.1
		SD	178.2	175.0*	182.6
		CV	0.392	0.364*	0.393
		RMSE	183.5	183.4*	190.4
		RRMSE	0.404	0.381*	0.408
	Soybeans	Bias	-1.03*	32.3	1.34
		RB	-0.084	-0.016*	-0.072
		AB	40.6	52.1	38.2*
		SD	125.4	111.3*	121.5
		CV	0.315	0.219*	0.271
		RMSE	136.1	129.3*	132.5
		RRMSE	1.797	1.286*	1.288

Table 3 shows for each state/crop combination the percent of replications where the simulated county-level hybrid estimators showed a positive bias (i.e., were higher than the corresponding population truth values). The bias statistics were computed over all estimated counties in the state combined. Note that the values for AEBP/RGS ranged from 66 to 75.1 percent, clearly suggesting a strong positive bias tendency for that estimator. By contrast, the bias values for BHF/RGS ranged from 43.7 to 51.9 percent and those for *BC* from 46.6 to 54.5 percent.

Table 3. Percent of Replications with Positive Bias (All Counties Combined)

State	**Crop**	**Hybrid Estimator**		
		BHF/RGS	AEBP/RGS	BC/RGS
Ohio	Corn	51.9	74.9	51.9
	Soybeans	48.6	66.3	51.7
South Dakota	Corn	50.4	75.1	54.5
	Soybeans	43.7	66.0	46.6

The three hybrid estimators were ranked from 1 to 3 by replication for each county based on absolute bias, with the ranks then averaged over all replications and counties. Table 4 displays these average ranks for each state/crop combination. Note that the average rank of BC/RGS was lowest in three of the four cases while that of AEBP/RGS was highest in all four cases.

Table 4. Average Ranks of Replication Level Absolute Bias (All Counties Combined)

State	**Crop**	**Hybrid Estimator**		
		BHF/RGS	AEBP/RGS	BC/RGS
Ohio	Corn	2.0	2.23	1.76
	Soybeans	1.87	2.23	1.88
South Dakota	Corn	2.07	2.2	1.73
	Soybeans	2.05	2.11	1.84

While Tables 2 through 4 show summary values of metrics averaged over counties, further insights are provided by Figures 1 through 12 which display box plots of three different metrics at the individual county level for each of the four state/crop combinations (bias in Figures 1-4, standard deviation in Figures 5-8 and relative root mean squared error in Figures 9-12). Note from Figures 1-4 that AEBP/RGS shows positive bias tendencies for all four state/crop combinations while the values for BHF/RGS and BC/RGS are more symmetric around zero. Examination of the SD and RRMSE plots (Figures 5-12) confirms the earlier observation of more favorable properties for AEBP/RGS than the other two hybrid estimators in terms of variability.

Figure 1. Box Plots of County-Level Bias for Corn in Ohio

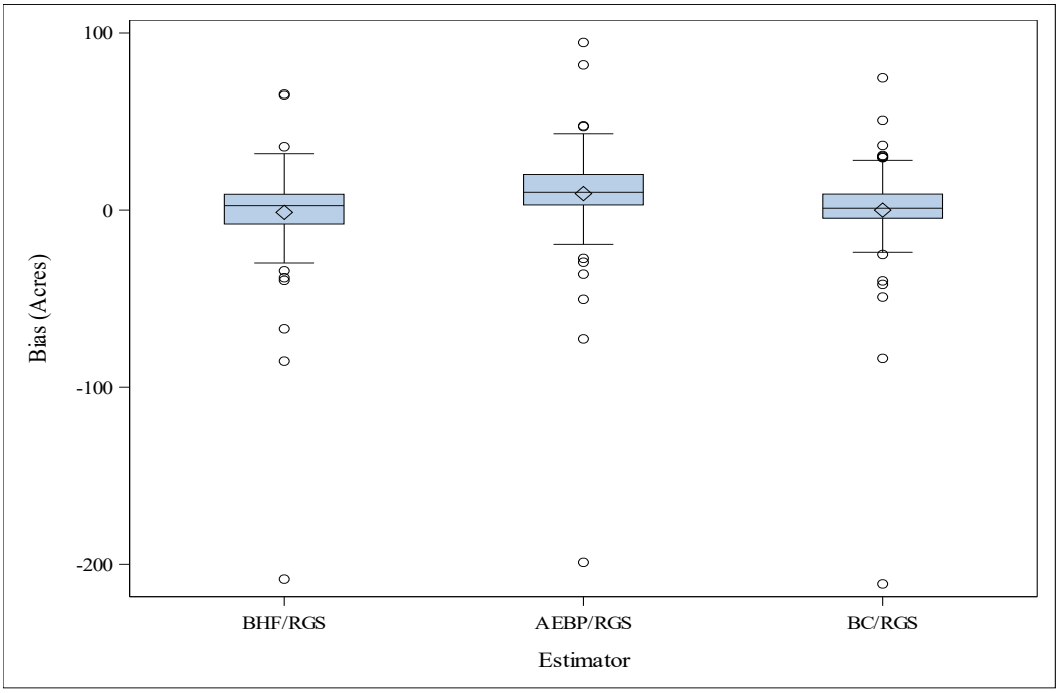


Figure 2. Box Plots of County-Level Bias for Soybeans in Ohio

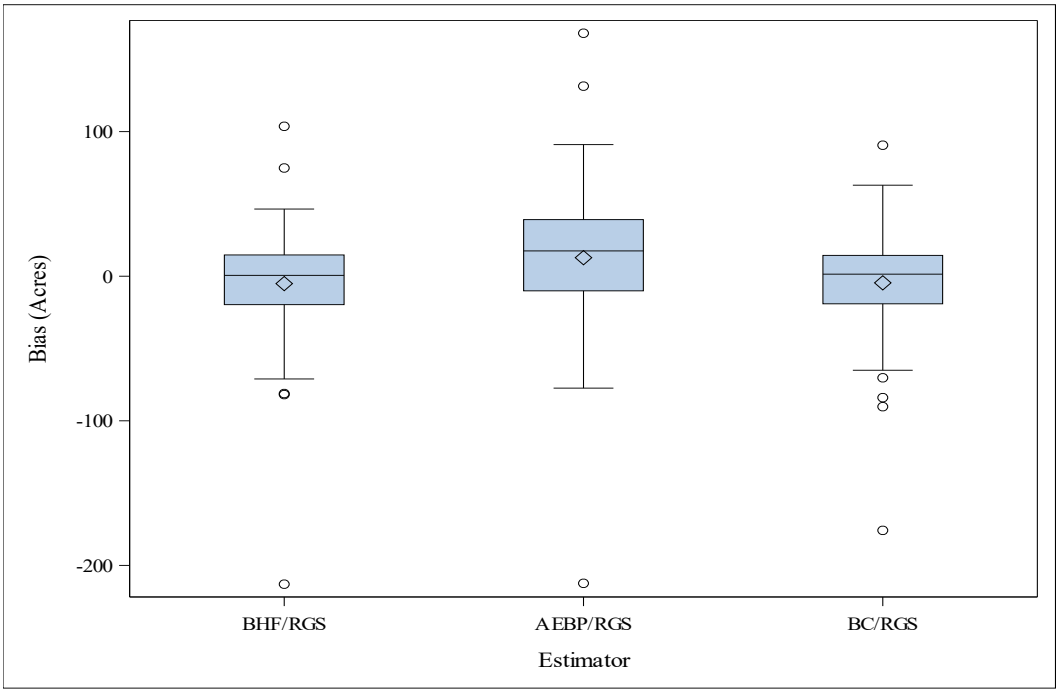


Figure 3. Box Plots of County-Level Bias for Corn in South Dakota

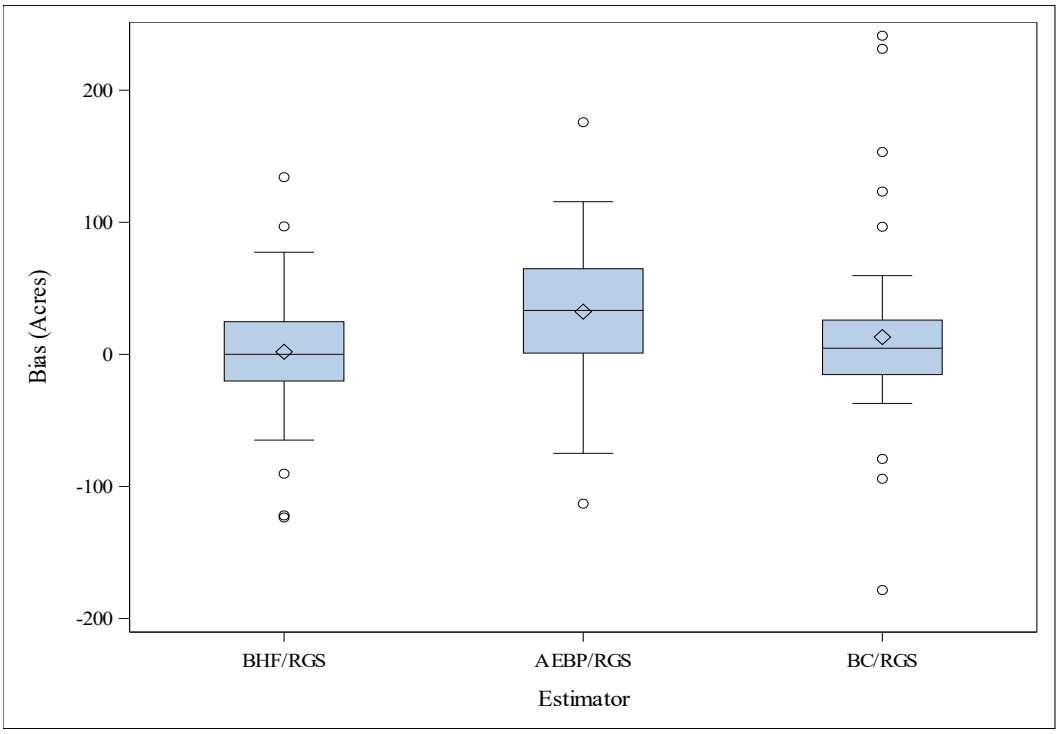


Figure 4. Box Plots of County-Level Bias for Soybeans in South Dakota

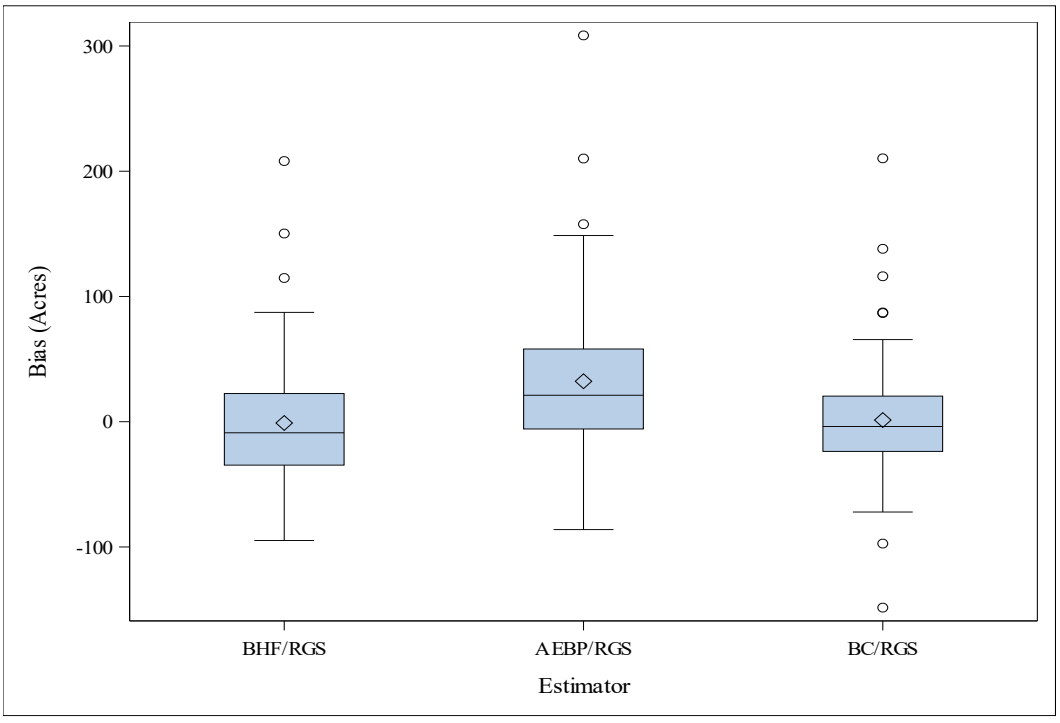


Figure 5. Box Plots of County-Level Standard Deviation for Corn in Ohio

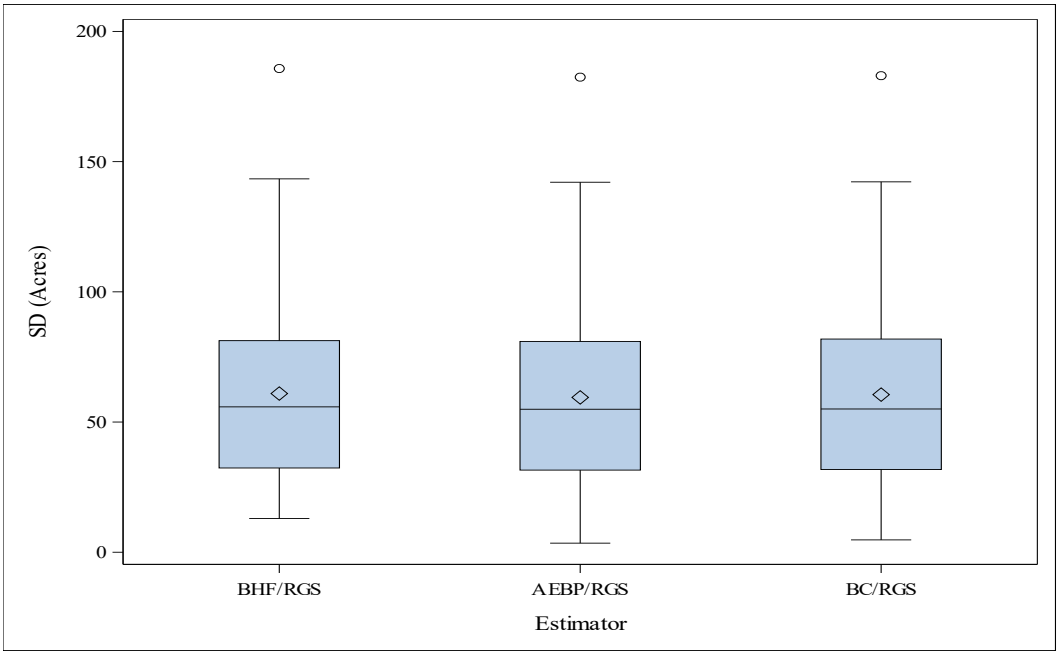


Figure 6. Box Plots of County-Level Standard Deviation for Soybeans in Ohio

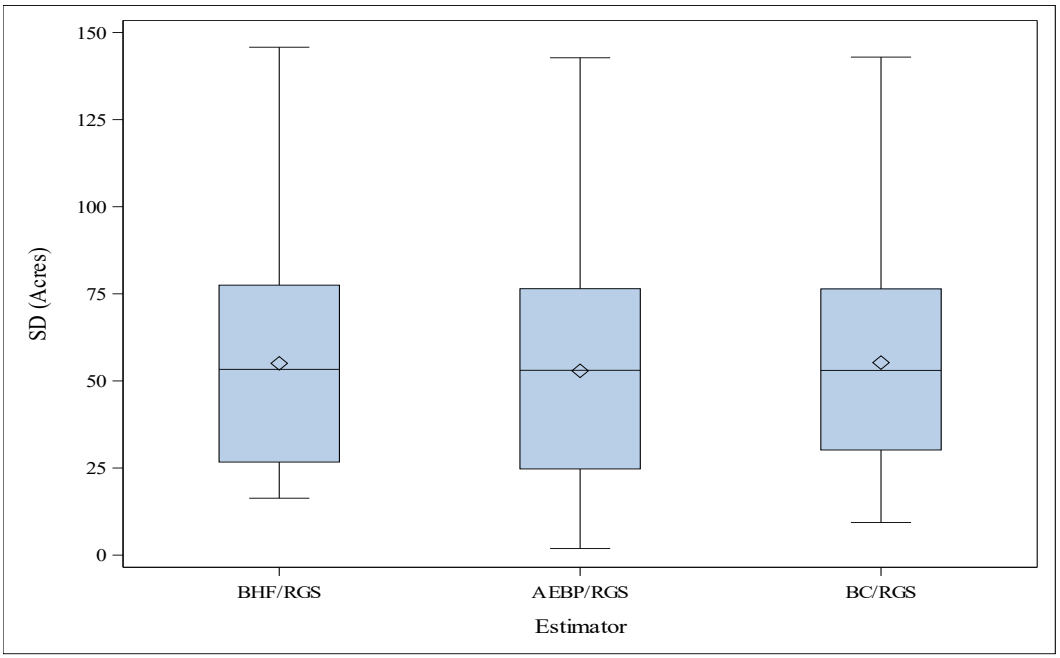


Figure 7. Box Plots of County-Level Standard Deviation for Corn in South Dakota

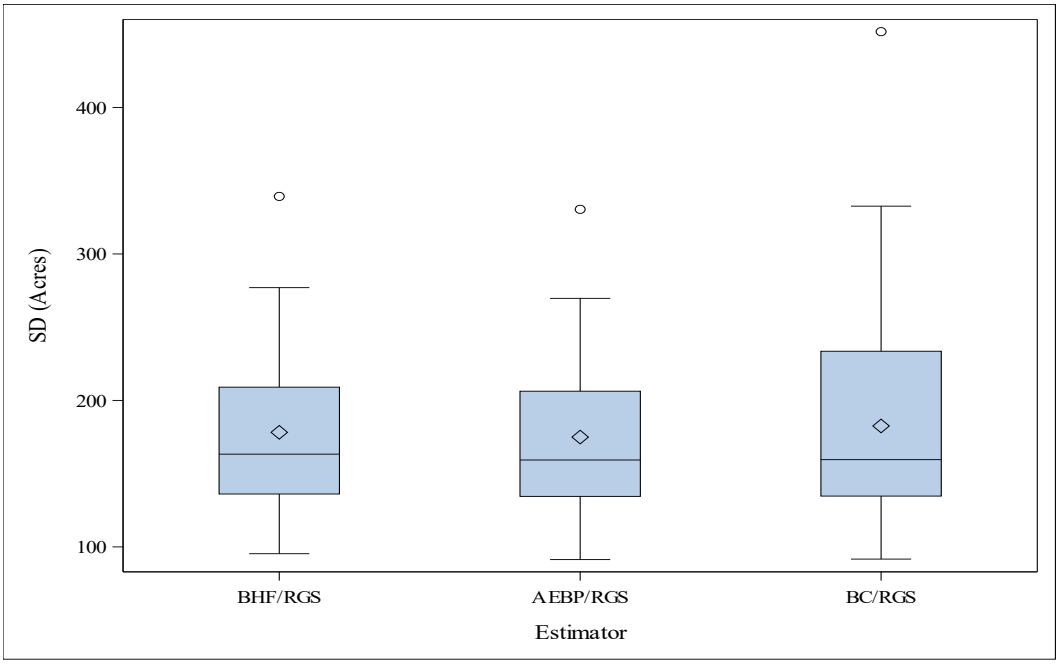


Figure 8. Box Plots of County-Level Standard Deviation for Soybeans in South Dakota

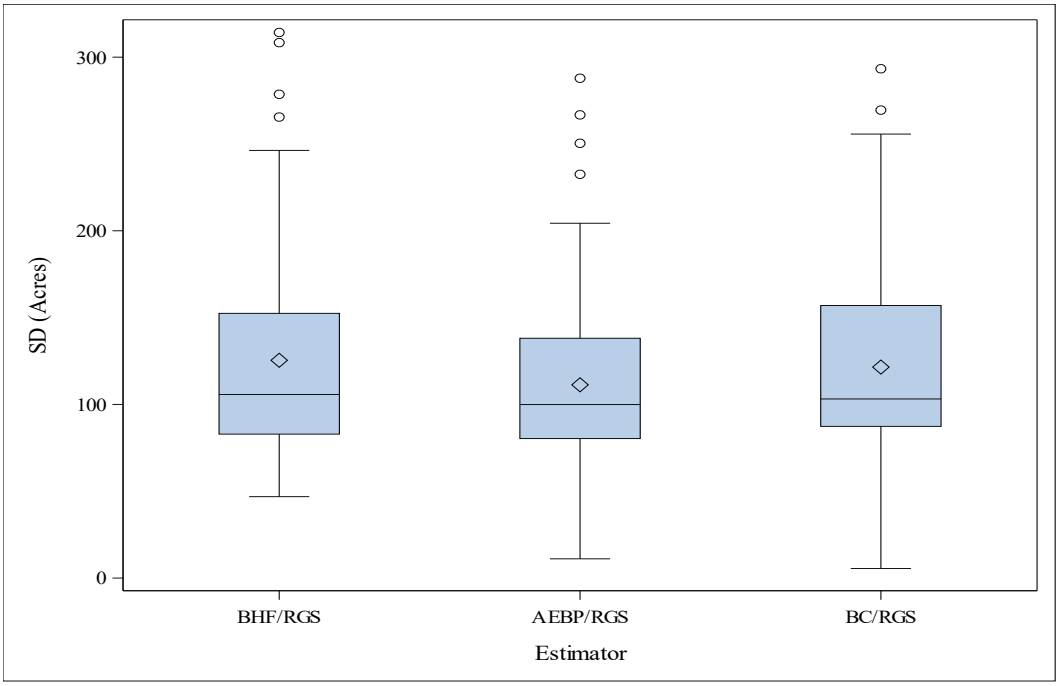


Figure 9. Box Plots of County-Level RRMSE for Corn in Ohio

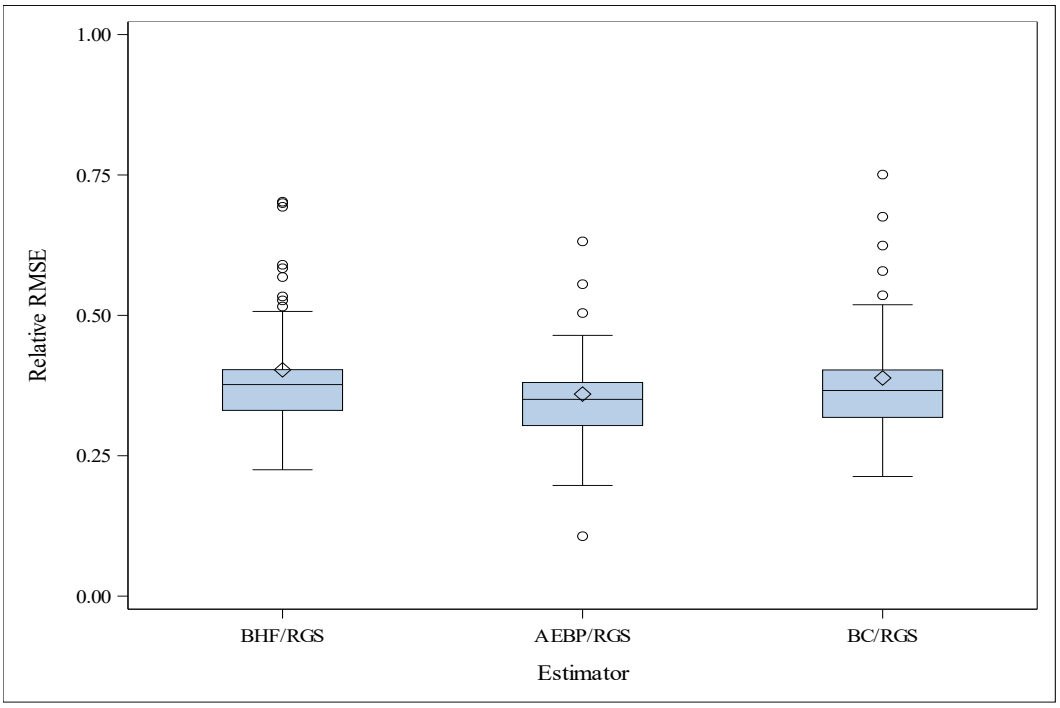


Figure 10. Box Plots of County-Level RRMSE for Soybeans in Ohio

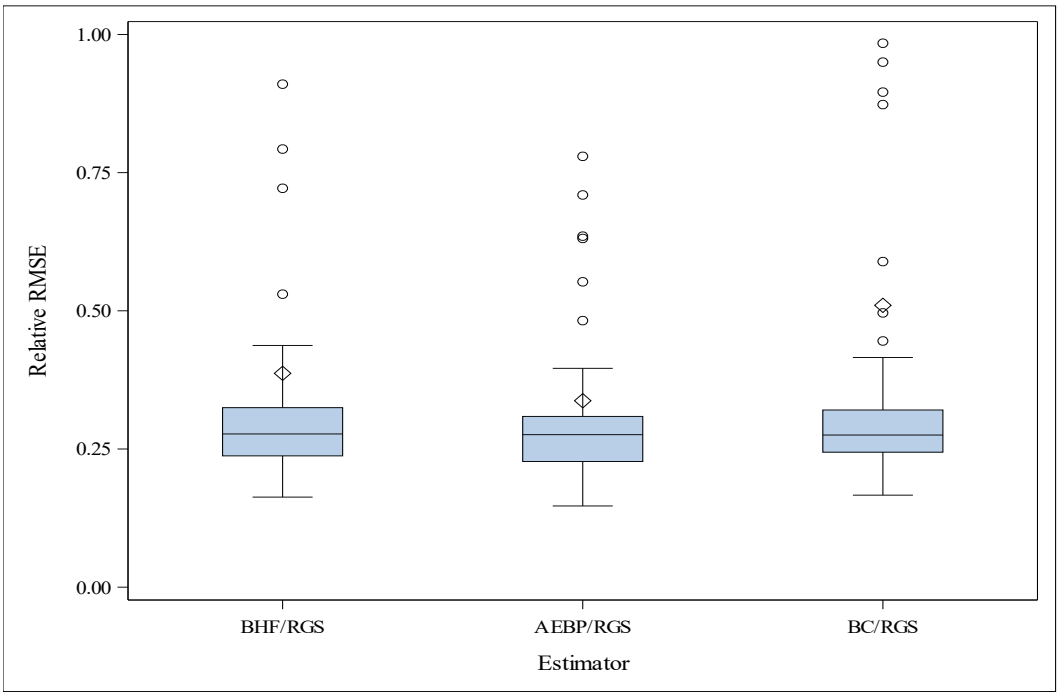


Figure 11. Box Plots of County-Level RRMSE for Corn in South Dakota

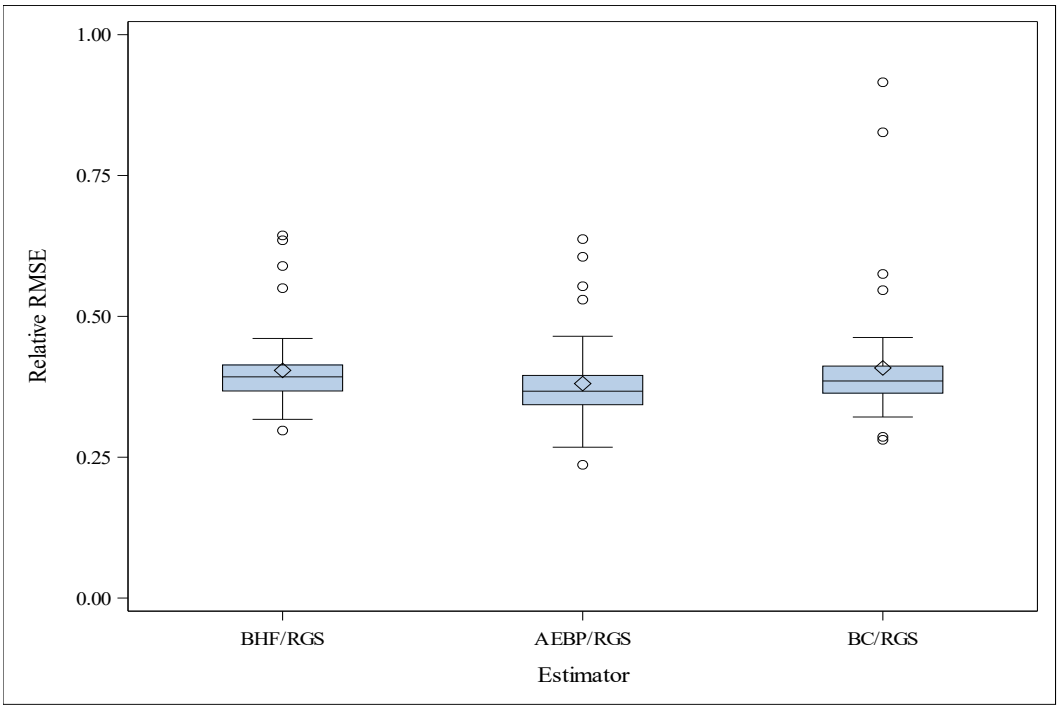
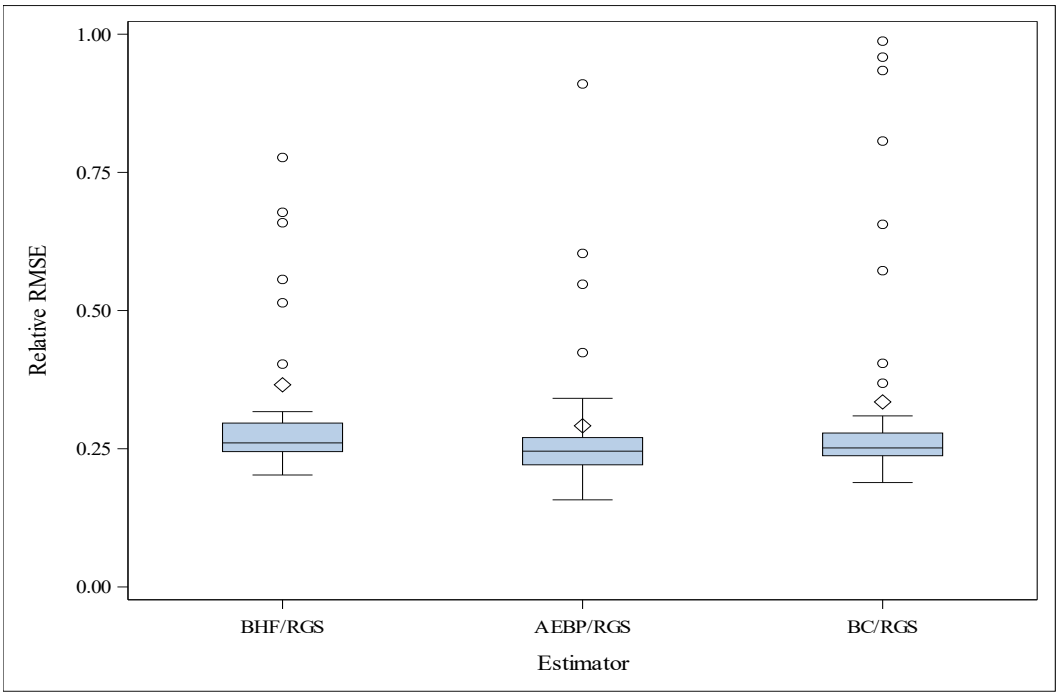


Figure 12. Box Plots of County-Level RRMSE for Soybeans in South Dakota



4. Summary and Future Work

Three mixed-model estimators for estimation of county-level crop planted area were evaluated via a simulation study involving corn and soybeans in two states (Ohio and South Dakota). The estimation procedures made use of combined County Agricultural Production Survey (CAPS) and auxiliary data, with a list frame size variable used as the unit (farm) level covariate and FSA planted acreage figures as the area (county) level covariate. Regression synthetic estimation was used for the subset of population units where the unit-level covariate was missing (so the actual estimators evaluated were of the hybrid variety).

The study found that the AEBP/RGS estimator displayed a strong tendency toward positive bias while BC/RGS appeared to be closest to unbiasedness among the three estimators. In terms of variability, AEBP/RGS displayed the most favorable properties among the three hybrid estimators evaluated.

Potential future directions for research would include a comprehensive comparison between the FSA and RS area-level covariates, further evaluation of the three mixed model estimators via simulation in other states and for additional crops, an investigation of robustness to departures from model assumptions and development of an adaptive version of the log-transformed Berg-Chandra estimator.

5. References

Battese, G.E., Harter, R.M. and Fuller, W.A. (1988). An Error-Components Model for Prediction of County Crop Areas using Survey and Satellite Data. *Journal of the American Statistical Association*, 83 (401), 28-36.

Bellow, M. and Lahiri, P. (2011). An Empirical Best Linear Unbiased Prediction Approach to Small Area Estimation of Crop Parameters. *Proceedings of the Joint Statistical Meetings, Section on Survey Research Methods,* 3976-3986.

Bellow, M. and Lahiri, P. (2012). Evaluation of Methods for County-Level Estimation of Crop Harvested Area that Employ Mixed Models. *Proceedings of ICES IV,* Montreal, Canada. https://ww2.amstat.org/meetings/ices/2012/papers/302087.pdf

Berg, E. and Chandra, H. (2014). Small Area Prediction for a Unit-Level Lognormal Model. *Computational Statistics and Data Analysis* Vol. 78, 159-175.

Boryan, C., Yang, Z., Mueller, R., Craig, M. (2011). Monitoring US Agriculture: the US Department of Agriculture, National Agricultural Statistics Service Cropland Data Layer Program, *Geocarto International*, Vol. 26, No. 5, 341-358. https://www.tandfonline.com/doi/abs/10.1080/10106049.2011.562309

Chatterjee, S., Lahiri, P. and Li, H. (2008). Parametric Bootstrap Approximation to the Distribution of EBLUP and Related Prediction Intervals in Linear Mixed Models. *Annals of Statistics*, Vol. 36, No. 3, 1221-1245.

Cruze, N., Erciulescu, A.L., Nandram, B., Barboza, W.J. and Young, L. (2019). Producing Official County-Level Agricultural Estimates in the United States: Needs and Challenges. *Statistical Science*, Vol. 34, No. 2, 301-316.

Jiang, J., and Lahiri, P. (2006). Mixed Model Prediction and Small Area Estimation, Editor's Invited Discussion Paper. *Test,* Vol. 15, 1-96.

Kott, P.S. and Bailey, J. (2000). The Theory and Practice of Maximal Brewer Selection with Poisson PRN Sampling. Invited Paper, *Proceedings of the Second International Conference on Establishment Surveys,* Buffalo, NY, 269-78.

Molina, I. and Rao, J.N.K. (2010). Small Area Estimation of Poverty Indicators. *Canadian Journal of Statistics*, Vol. 38, No. 3, 369-385.

Pfeffermann, D. Terryn, B. and Moura, F.A.S. (2008). Small Area Estimation Under a Two-Part Random Effects Model with Application to Estimation of Literacy in Developing Countries, *Survey Methodology*, Vol. 34, No. 2, 235-249.