An Empirical Best Linear Unbiased Prediction Approach to Small-Area Estimation of Crop Parameters

Michael E. Bellow and Partha S. Lahiri
National Agricultural Statistics Service, 3251 Old Lee Hwy., Fairfax, VA 22030
JPSM, University of Maryland, 1218 LeFrak Hall, College Park, MD 20742

Abstract

Accurate county (small-area) level estimation of crop and livestock items is an important priority for the USDA's National Agricultural Statistics Service (NASS). We consider an empirical best linear unbiased prediction (EBLUP) method for combining multiple data sources to estimate crop harvested area (and potentially other crop parameters) at the county level. This method assumes a linear mixed model that relates survey reported harvested area to both unit (farm) and area (county) level covariates, with variance components estimated using a technique which ensures strictly positive consistent estimation of the model variance. A parametric bootstrap method that incorporates all sources of uncertainty can be used to estimate variability parameters. Results of a study comparing the proposed EBLUP method with standard ratio and regression type estimators and a synthetic estimator for corn and soybeans in seven states in the Midwestern grain belt region of the US are discussed.

Key Words: Small-Area Estimation, Components of Variance, Predictor Variables

1. Introduction

Various government agencies (e.g., the United States Census Bureau, USDA's National Agricultural Statistics Service (NASS), Statistics Canada and the Central Statistical Office of the United Kingdom) have a requirement to produce reliable small-area statistics. A small-area generally refers to a geographical entity (e.g., US county) for which limited information is available from the primary source of data. Accurate small-area statistics are needed for regional planning and fund allocation in many government programs and thus their importance cannot be overemphasized. County estimates of crop parameters such as harvested area, production and yield are used by farmers, agribusinesses and government agencies for local agricultural decision making.

NASS has been publishing county level crop and livestock inventories since 1917 (see Iwig (1993) for historical background). The main source of data used by the agency for commodity estimation has always been its surveys of farmers, ranchers and agribusiness managers who provide requested information on a voluntary, confidential basis. Since surveys designed and conducted at the national and state levels are seldom adequate for obtaining reliable county level estimates, NASS has made extensive use of ancillary data sources such as list sampling frame control data, previous year estimates, earth observing satellite data and census of agriculture data in its county estimation procedures.

NASS uses data from multiple sample surveys to estimate harvested area, yield and production for various crops at the county level. The estimates of harvested area and production are constrained to sum to NASS official district (subdivision of a state into

agricultural regions) and state level totals, with yield estimates required to be consistent with such totals. The census of agriculture (conducted twice each decade by NASS in years ending in '2' and '7') serves as a useful benchmark for county estimation.

In general, traditional direct methods that utilize only small-area specific survey data are highly unreliable, mainly due to small sample sizes in the areas of interest. In addition, effects of nonsampling errors such as coverage and nonresponse can be severe, and combining data from several different surveys doesn't often resolve this problem satisfactorily. In order to improve on direct estimation, several indirect and model-based methods have been proposed. These procedures essentially use implicit or explicit models that borrow strength from related resources such as administrative and census records as well as survey data from earlier years. Rao (2003) and Jiang and Lahiri (2006) provide comprehensive reviews of different small-area estimation methods and applications.

Previously, we evaluated an empirical Bayes approach to estimation of crop harvested yield (Bellow and Lahiri, 2010). Although this paper focuses specifically on estimation of harvested area, the methodology could potentially be extended to other crop related items such as production. Section 2 introduces an empirical best linear unbiased predictor (EBLUP) estimator that relates a dependent variable to both unit (farm) and area (county) level auxiliary (predictor) variables. Section 3 discusses results of a study evaluating the efficiency of the EBLUP method (using one farm level and one county level covariate) for estimation of corn and soybean harvested acreage at the county level in seven Midwestern states for 2008.

2. EBLUP Method for Harvested Area Estimation

This section describes a proposed empirical best linear unbiased prediction. Before proceeding, we need the following notation:

L = number of counties to be estimated (in state),

U_i = target population of all farms in county *i* (*i*=1,...,L),

y_{ij} = survey reported harvested area of crop of interest in county *i*, farm *j*, and

$Y_i = \sum_{j \in Ui} y_{ij}$ (total harvested area of crop in county *i*).

Note that $y_{ij} = 0$ if farm *j* of county *i* (in the list frame) did not grow the crop of interest for the year under consideration and $y_{ij} > 0$ otherwise. Thus we have:

$$Y_i = \sum_{j \in Ui+} y_{ij}$$

where:

U_{i+} = subset of U_i containing farms with strictly positive harvested area.

Since sample sizes within counties area generally small, any direct estimator of the county level population totals would be unreliable. Therefore we introduce an empirical Bayes method that borrows strength from a set of auxiliary variables available at either the farm or county level.

We consider a components of variance model based on the one developed by Battese, Harter and Fuller (1988) under a cooperative agreement between Iowa State University and NASS (for an application involving the use of Landsat data to improve the efficiency of county level crop planted area estimates). The model (with *p* auxiliary variables) is:

$$yij = x_{ij}^T \beta + vi + eij$$

where:

y_{ij} = survey reported harvested acreage in county *i*, sampled farm *j*,

$x_{ij} = (x_{1ij} \quad x_{2ij} \quad \ldots \quad x_{pij})^T$,

x_{kij} = value of *k*th auxiliary variable for county *i*, sampled farm *j*,

$= (\quad _1 \quad _2 \quad \ldots \quad _p)^T$,

$_k$ = regression parameter associated with auxiliary variable *k,*

v_i = county effect for county *i*, and

e_{ij} = random error term.

The farm level auxiliary variables are assumed to be available for every population unit in the counties being estimated. The county effects $\{v_i\}$ attempt to capture the variation not accounted for by the auxiliary variables and are assumed to be independent $N[0, \ ^2_v]$ random variables. The $\{e_{ij}\}$ are assumed to be independent $N[0, \ ^2_e]$ random variables and independent of the $\{v_i\}$.

The estimation procedure is as follows:

1) Fit non-missing survey reported harvested acreage values $\{y_{ij}\}$ to the auxiliary variables $\{x_{kij}\}$ (k=1,...,p) using the restricted maximum likelihood (REML) method, thereby obtaining estimates of the regression parameters and county effects.

2) Compute estimate of total harvested acreage in each county *i* as:

$$\hat{Y}_i^{(EB)} = N_i[\hat{\beta}_0 + \sum_{k=1}^{p} \hat{\beta}_k \bar{X}_{ki} + \hat{v}_i]$$

where:

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

$\bar{X}_{ki}$ = population mean of auxiliary variable *k* in county *i* (*k*=1,...,*p*; *i*=1,...,*L*),

$\hat{\ }_k$ = REML estimate of $_k$ (*k*=0,...,*p*), and

$\hat{v}_i$ = REML estimate of v_i (*i*=1,...,*L*).

For county level auxiliary variables, the population means in a given county are assumed to be the respective values of those variables for that county.

The above procedure assumes that values for the farm level auxiliary variables are available for every population unit in the sampling frame. If there are population units with a missing value for at least one farm level covariate, they must be excluded from the computations. The resulting EBLUP estimates for counties containing such population units will thus be based on less than full coverage of the population. For that reason, another type of estimation (such as synthetic) should be applied in the portion of the frame corresponding to those units to compensate for the undercoverage. EBLUP estimates for counties with a missing value for one or more of the county level covariates cannot be computed.

As a final step, the EBLUP estimates (after possibly being combined with synthetic or other estimates to compensate for undercoverage) are benchmarked for consistency with NASS official state level estimates of harvested area. Estimation of variability parameters of the EBLUP estimator (such as mean squared error) can be done using a parametric bootstrap method along the lines of Chatterjee, Lahiri and Li (2008).

3. Results of Empirical Study

The proposed EBLUP estimator of harvested area was evaluated for corn and soybeans in the 2008 crop season. The study area included the following seven states in the Midwestern grain belt region of the United States: 1) Illinois, 2) Indiana, 3) Iowa, 4) Kansas, 5) Minnesota, 6) Missouri, and 7) Ohio. The two auxiliary variables used were: 1) list frame size variable (farm level), and 2) Farm Service Agency (FSA) county level planted acreage for 2008.

NASS's list sampling frame (for a given state) is a periodically updated list of farm names and addresses along with control data that identify the relative size of the items being surveyed. The farm level auxiliary variable used in the EBLUP model was the previous year's (2007) *size variable,* a measure of planted acreage for the crop of interest obtained

from the previous year's list frame. The size variable is derived as the maximum reported planted acreage for a farm over a prespecified number of years and can be zero, positive or missing. A missing value for the size variable could arise if the farm in question is in the list frame for the current year but not the previous year, while a zero value could occur if the crop of interest was never planted on the farm over the time period used in its computation.

The second auxiliary variable consisted of planted acreage estimates at the county level released by USDA's Farm Service Agency. These figures are derived from mandatory reports submitted by farmers to local FSA offices and are generally available to NASS in time for use in its own operational county estimation.

The list frame units are divided into groups based on the value of the size variable, with one of the groups containing all units with a missing value for the size variable. For purposes of this paper, the groups corresponding to non-missing values of the size variable are treated as a single entity (called stratum *A*) with the 'missing' group referred to as stratum *B*. Since for obvious reasons the EBLUP estimation procedure could only be done in stratum *A*, synthetic (SYN) estimation was used for stratum *B*. The synthetic estimates were computed as follows:

$$\hat{Y}_{Bi}^{(SYN)} = N_{Bi}\,\bar{y}_B$$

where:

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

$\bar{y}_B$ = sample mean reported harvested acreage in stratum *B* (over all counties)

Tables 1 and 2 give estimated values of the parameters obtained from fitting the EBLUP model in each of the seven states for corn and soybeans, as well as *p*-values associated with Student's *t*-tests of whether $_1$ and $_2$ are significantly different from zero. Table 1 shows that the farm level size variable was significant at the 0.0001 level for corn in all seven states, while the county level FSA planted acreage variable was significant at the 0.001 level in six of the states and at the 0.05 level in all seven. From Table 2, the farm level variable was significant at the 0.001 level for soybeans in all seven states but the county level variable was significant at the 0.05 level in only three states (Kansas, Missouri and Ohio).

For each state, the EBLUP estimator was compared with four survey estimators commonly used when population level auxiliary information is available - 1) the simple ratio estimator (SR), 2) combined ratio estimator (CRE), 3) simple regression estimator (SRGE), and 4) combined regression estimator (CRGE) (Cochran, 1977). The pure synthetic estimator (SYN) computed by applying equation (1) for strata A and B combined (instead of just stratum B) was also used in the estimator comparison. Official NASS county level harvested acreage estimates for 2008 were used as the 'gold standard' for assessing estimation accuracy in the study.

Five efficiency metrics were computed for each of the five estimate types - average absolute deviation (AAD), average squared deviation (ASD), average absolute relative deviation (AARD), average squared relative deviation (ASRD) and percentage below official (PBO). AAD is computed as the mean of absolute deviations between county estimates and

corresponding 2007 official estimates, ASD the corresponding mean of squared deviations, AARD the mean of ratios between absolute deviations and official values and ASRD the corresponding mean of squared ratios. PBO is defined as the proportion of counties with estimate less than the corresponding 2008 official estimate. Values of PBO below (above) 0.5 suggest overestimation (underestimation) tendencies for an estimator. Note that while the first four metrics shed light on magnitude of bias and variability, PBO is the only one of the five that relates to direction of bias

Table 1. Estimated EBLUP Model Parameters for Corn

State	Intercept	Auxiliary Variable 1		Auxiliary Variable 2	
	$\hat{\beta}_0$	$\hat{\beta}_1$	*p*-value	$\hat{\beta}_2$	*p*-value
Illinois	-42.7	0.94	<0.0001	0.0003	<0.0001
Indiana	-13.5	0.8	<0.0001	0.0006	<0.0001
Iowa	5.9	0.81	<0.0001	0.0002	0.0005
Kansas	1.4	0.86	<0.0001	0.0005	0.037
Minnesota	-21.2	0.87	<0.0001	0.0002	<0.0001
Missouri	-14.8	0.72	<0.0001	0.0006	<0.0001
Ohio	-6.1	0.8	<0.0001	0.0002	0.001

Table 2. Estimated EBLUP Model Parameters for Soybeans

State	Intercept	Auxiliary Variable 1		Auxiliary Variable 2	
	$\hat{\beta}_0$	$\hat{\beta}_1$	*p*-value	$\hat{\beta}_2$	*p*-value
Illinois	6.1	0.94	<0.0001	0.00004	0.67
Indiana	4.5	0.94	<0.0001	0.0002	0.08
Iowa	8.9	0.89	<0.0001	0.00001	0.92
Kansas	6.5	0.94	<0.0001	0.0005	0.007
Minnesota	3.5	0.91	<0.0001	0.00006	0.23
Missouri	-28.1	0.88	<0.0001	0.0007	<0.0001
Ohio	9.5	0.88	<0.0001	0.0001	0.04

Figure 1. Comparison of Average Absolute Relative Deviation for Corn

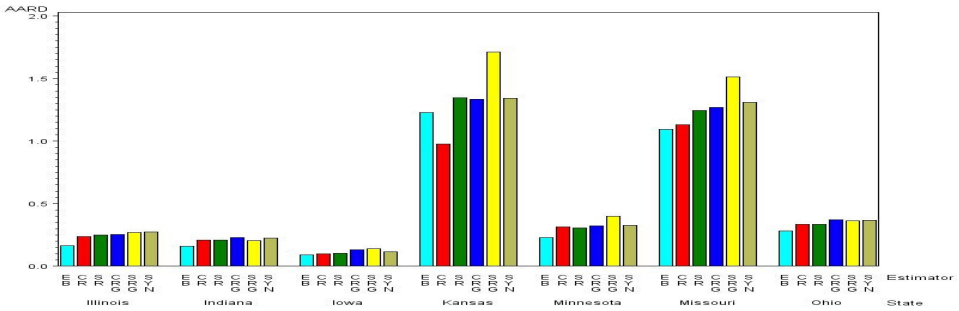


Figure 2. Comparison of Percentage Below Official for Corn

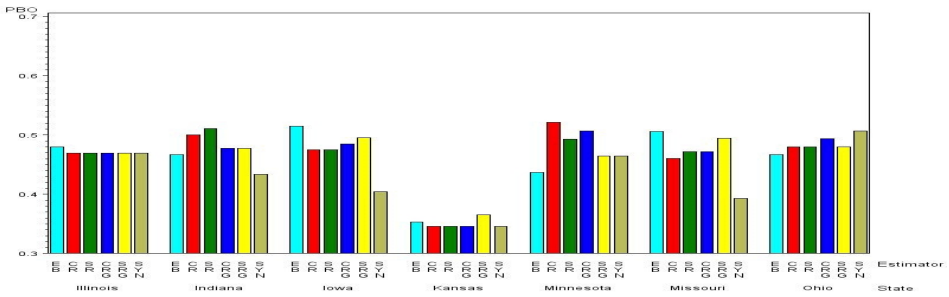


Figure 3. Comparison of Average Absolute Relative Deviation for Soybeans

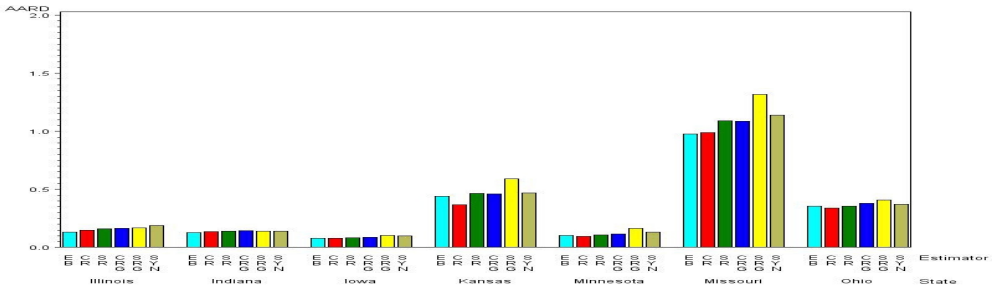
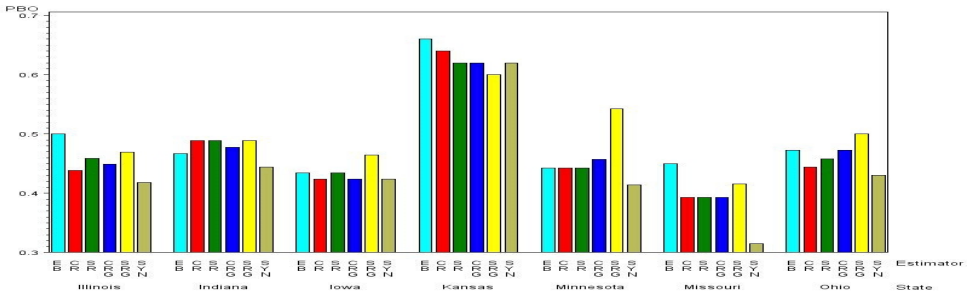


Figure 4. Comparison of Percentage Below Official for Soybeans



Figures 1 through 4 are bar charts displaying the computed values of two of the metrics (AARD and PBO) for corn and soybeans, respectively. Figures 1 and 3 show notably high values of AARD for all six estimators in Kansas and Missouri compared with the other five states. Figure 2 indicates strong positive bias tendencies with all of the estimators for corn in Kansas, while Figure 4 shows negative bias tendencies for soybeans in the same state. Note also that EBLUP had the lowest AARD among the six estimators for six of the seven states for corn (with Kansas being the exception), but only three states (Illinois, Indiana and Missouri) for soybeans.

Table 3. Average Estimator Ranks by State for Corn

State	**Estimator**					
	EBLUP	**CR**	**SR**	**CRG**	**SRG**	**SYN**
Illinois	1.0*	2.4	3.2	4.2	4.8	5.4
Indiana	1.8*	3.0	3.2	5.5	2.3	5.2
Iowa	1.3*	2.5	3.3	4.9	4.6	4.4
Kansas	2.0	1.7*	4.1	4.1	5.0	4.1
Minnesota	2.0	3.2	1.9*	3.7	4.9	5.3
Missouri	1.3*	2.4	3.1	3.9	4.7	5.6
Ohio	2.0*	3.0	2.8	4.4	4.2	4.6
All	**1.6***	**2.6**	**3.1**	**4.4**	**4.4**	**4.9**

* - best for state (or overall)

Table 4. Average Estimator Ranks by Metric for Corn

Metric	**Estimator**					
	EBLUP	**CR**	**SR**	**CRG**	**SRG**	**SYN**
AAD	1.1*	2.4	2.7	5.0	4.4	5.3
ASD	1.1*	2.7	2.7	5.0	4.1	5.3
AARD	1.1*	2.1	3.3	4.6	5.0	4.9
ASRD	1.3*	2.0	3.3	4.4	5.4	4.6
PBO	3.4	3.7	3.4	2.9	2.8*	4.7

* - best for state (or overall)

To further quantify the performance of EBLUP relative to the other five estimators, a rank based comparison was done. The estimators were ranked from 1 (best) to 6 (worst) based on each of the five metrics, with the ranks for PBO computed based on absolute deviation from 0.5, with the ranks average by state (over metrics) and by metric (over states) and overall. Tables 3 and 4 show the average ranks by state and overall for corn and soybeans, respectively.

Table 5. Average Estimator Ranks by State for Soybeans

State	**Estimator**					
	EBLUP	**CR**	**SR**	**CRG**	**SRG**	**SYN**
Illinois	1.2*	2.4	3.2	4.2	4.2	5.8
Indiana	1.8*	2.4	3.4	5.4	4.2	3.8
Iowa	1.3*	2.6	2.9	4.2	4.6	5.4
Kansas	2.8	1.8*	3.8	3.0	5.0	4.6
Minnesota	2.4	1.6*	3.2	3.6	5.0	5.2
Missouri	1.2*	2.2	3.8	3.6	4.8	5.4
Ohio	3.3	1.8*	3.0	4.1	3.6	5.2
All	**2.0***	**2.1**	**3.3**	**4.0**	**4.5**	**5.1**

* - best for state (or overall)

Table 6. Average Estimator Ranks by Metric for Soybeans

Metric	**Estimator**					
	EBLUP	**CR**	**SR**	**CRG**	**SRG**	**SYN**
AAD	1.9	1.7*	3.1	4.3	4.9	5.1
ASD	1.9	1.7*	3.3	4.3	4.7	5.1
AARD	1.4*	1.6	3.4	4.1	5.7	4.7
ASRD	1.7	1.3*	3.6	3.9	5.7	4.9
PBO	3.1	4.3	3.2	3.5	1.4*	5.4

* - best for state (or overall)

Table 3 shows that the EBLUP estimator had lower average rank (over the five metrics) for five of the seven states and overall (all states combined) for corn. The two exceptions were Kansas and Minnesota where either the combined ratio estimator or the separate ratio estimator had a slightly lower average rank than EBLUP. Table 4 indicates that EBLUP had

the lowest average rank for four of the five metrics for corn, with the exception being PBO for which the separate regression and combined regression estimators were better.

The results for soybeans were somewhat less favorable to the EBLUP estimator. Table 5 shows it having lowest average rank in four of the seven states (and overall), with the exceptions being Kansas, Minnesota and Ohio where the combined ratio estimator was better. Table 6 shows that EBLUP had lowest average rank for only one metric (AARD), although it was only slightly worse than CR for AAD and ASD and uniformly better than the other four estimators for all metrics other than PBO (for which the separate regression estimator had the lowest average rank).

4. Summary and Comments

An empirical best linear unbiased prediction (EBLUP) approach to estimation of crop parameters such as harvested area was proposed. This method uses auxiliary information available both at the population unit (farm) level and area (county) level. The EBLUP estimator was tested for corn and soybeans in 2008 using two auxiliary variables: 1) farm level size variable, and 2) county level FSA planted acreage figures. Synthetic estimation was used to adjust for undercoverage due to missing values of the first covariate.

Based on five metrics, the EBLUP estimator was found to outperform the other five estimators in general for both crops (especially corn). The gains in efficiency would probably have been more pronounced had there been no population units with missing data for the farm level size variable so that a pure EBLUP estimator could have been used (without needing to combine it with a synthetic estimator). Future research will focus on possible modifications and refinements to the EBLUP methodology described in this paper, with special attention to a modelling approach involving log-transformed variables.

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, Vol. 83, 28-36.

Bellow, M. and Lahiri, P. (2010), "Empirical Bayes Methodology for the NASS County Estimation Program." Proceedings of the Section on Survey Research Methods, American Statistical Association, 343-355.

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.

Cochran, W.G. (1977), *Sampling Techniques*. New York, NY: John Wiley & Sons, Inc.

Iwig, W.C. (1993), "The National Agricultural Statistics Service County Estimates Program." In *Indirect Estimators in Federal Programs*, Statistical Policy Working Paper 21, Subcommittee on Small Area Estimation, Federal Committee on Statistical Methodology, Office of Management and Budget.

Jiang, J., and Lahiri, P. (2006), "Mixed Model Prediction and Small Area Estimation (with Discussions)." TEST: An Official Journal of the Spanish Society of Statistics and Operations Research, Vol. 15, Issue 1, 1-96.

Rao, J. N. K., (2003), *Small Area Estimation*. New York, NY: John Wiley & Sons, Inc.