Model-Based County-Level Crop Estimates Incorporating Auxiliary Sources of Information

Andreea L. Erciulescu[*] Nathan B. Cruze [†] Balgobin Nandram[‡]

Abstract
In 2011, USDA's National Agricultural Statistics Service started the complete implementation of the County Agricultural Production Survey (CAPS). CAPS is an annual survey to provide accurate county-level acreage and production estimates of approved federal and state crop commodities. The current top-down method of producing official county-level estimates that satisfy the county-district-state benchmarking constraint is an expert assessment incorporating multiple sources of information. We propose a model-based method that combines the CAPS survey acreage data with auxiliary data and improves county-level survey estimation, while providing measures of uncertainty for the county-level acreage estimates. Auxiliary sources of information include remote sensing, weather data, and planted acreage administrative data from other USDA agencies. A novel hierarchical Bayesian subarea-level model is proposed and implemented, with an additional hierarchical level for the sampling variances. County-level model-based acreage estimates have lower coefficients of variation than the corresponding county-level survey acreage estimates. Top-down benchmarking methods are investigated and the final acreage estimates satisfy the county-district-state benchmarking constraint.

Key Words: Auxiliary Data, Benchmarking, Crop Acreage Estimates, Hierarchical Bayes, Small Area Estimation.

1. Introduction

The USDA's National Agricultural Statistics Service (NASS) county-level estimates of acreage, production and yield may contribute to the magnitude of payout in some agricultural programs. Two major USDA agencies that use NASS's county-level estimates of acreage, production and yield for decision making are the Farm Service Agency (FSA) and the Risk Management Agency (RMA). Given their importance to USDA's mission and to the agricultural sector at large, it is important that NASS releases reliable county-level crop estimates.

NASS's quarterly Acreage, Production and Stocks (APS) surveys are designed to support national and state crop estimates released in annual summary reports (USDA NASS, 2016a). The County Agricultural Production Survey (CAPS) is an annual survey constructed to supplement NASS's quarterly APS surveys (which are designed for higher levels of aggregation) and provide data for more reliable acreage, production and yield county-level estimates of approved federal and state crop commodities. In 2011, USDA's NASS fully implemented the CAPS, for which the data collection window extends beyond the release of official state estimates. Therefore, state totals are published prior to county-level estimation.

[*] National Institute of Statistical Sciences and USDA National Agricultural Statistics Service, 1400 Independence Avenue, SW, Room 6050, Washington, DC 20250-2054. E-mail: aerciulescu@niss.org

[†] USDA National Agricultural Statistics Service, 1400 Independence Avenue, SW, Room 6412 A, Washington, DC 20250-2054

[‡] Worcester Polytechnic Institute and USDA National Agricultural Statistics Service, Department of Mathematical Sciences, Stratton Hall, 100 Institute Road, Worcester, MA 01609

Following NASS terminology, the estimates constructed using survey data or using auxiliary data are denoted by *indications*, while the official, published, estimates are denoted by *estimates* (Adrian, 2012). Survey indications are produced based on the direct expansion methods, and variances for the survey indications are computed using a delete-a-group Jackknife, with 15 replicate groups (Kott, 1998). Estimates are required at two hierarchical, substate levels: the county and the agricultural statistics district (ASD), which is a predefined group of neighboring counties.

Currently, NASS's published county-level estimates are produced using a top-down approach. The NASS Agricultural Statistics Board (ASB) sets the state-level estimates using the NASS's quarterly APS surveys information, as well as auxiliary data sources. Approximately one month after the release of state-level estimates, CAPS data collection concludes and ASB's substate-level estimation process begins. Using the combined APS and CAPS samples, as well as diverse sources of auxiliary information, the ASB sets the agricultural statistics district-level estimates to aggregate to the state-level estimates, and the ASB sets the county-level estimates to satisfy the county-district-state benchmarking constraint. Because the current expert assessment method of constructing the county-level estimates is not model-based, uncertainty measures for the official estimates are not available for publication. Moreover, NASS's county-level estimates are publicly available only for counties that meet the publication standards; for example, due to a small county-level sample size, approximately 64% of the corn county-level acreage estimates were published in 2014. We propose a model-based estimation approach that incorporates multiple sources of information to produce county-level estimates and associated measures of uncertainty for all the sampled counties. While the long-term goal is to develop a model-based estimation approach for county-level estimates of acreage, production and yield, in this paper we illustrate the methods for acreage only.

County-level survey indications may be improved using auxiliary information and small area model-based procedures. In this section we mention the pioneer research in small area model-based estimation, as well as previous NASS model-based application studies. Depending on the availability of the data sources, unit-level models, area-level models and subarea-level models are discussed. Rao and Molina (2015) provide an extended outline of small area estimation methods, including benchmarking techniques.

Fuller and Battese (1973) first developed the framework of nested error linear regression models, a methodology that is currently used by NASS to estimate survey reported crop planted acreage for an area frame sample unit, for selected commodities, using survey data and counts of classified pixels. For details on the current procedure, see Bellow (1993). The resulting model-based planted acreage indications represent one of the multiple sources of auxiliary data used in the current NASS county-level estimation process, described above. Battese, Harter and Fuller (1988) introduced the unit-level models for small area estimation based on nested error linear regression. The authors applied the estimation methods to the county crop area estimation, using survey and satellite data. Bellow and Lahiri (2011, 2012) propose county-level model-based estimates for crop harvested acreage, using extensions of the model in Battese, Harter and Fuller (1988). The unit-level response information is available from the CAPS and auxiliary information includes the NASS list sampling frame, administrative data, satellite data and Census of Agriculture data.

Fay and Herriot (1979) introduced the area-level models, popularizing model-based small area estimation methods. These models are excellent tools for summary survey data where the area-level survey indications and the area-level sampling variances are known. Simple estimation methods could be applied, under certain model assumptions, and the confidentiality of the unit-level data is protected. Bellow and Lahiri (2010) propose county-level model-based estimates for crop harvested yield, using the model in Fay and Herriot (1979). The survey indications and the sampling variances are available from the CAPS and auxiliary information includes the official NASS county-level crop production estimates and the Census of Agriculture production statistics.

While the unit-level models and the area-level models are useful tools in producing reliable area-level estimates, the hierarchical structure of the data and the consistency between the estimates for different hierarchical levels may not hold. More specifically, for a specific state composed of districts, where the districts are composed of counties, it is desirable that the county-level acreage estimates sum to the district-level acreage estimates and that the district-level acreage estimates sum to the state-level acreage estimate. Model-based estimates that account for the hierarchical structure of the data and that benefit from an automatic benchmarking to a higher level were first introduced by Fuller and Goyeneche (1998), in the context of Small Area Income and Poverty Estimation in the United States (Census 2016). The authors proposed a subarea-level model for an application where the subarea was the county, nested within a state (area). For further details on a similar subarea-level model see Torabi and Rao (2014) and Rao and Molina (2015).

In this paper we extend the model proposed by Fuller and Goyeneche (1998) to construct model-based county-level harvested acreage estimates. The smallest unit considered is the county, and the district-level benchmarking constraint is an implicit effect of the model proposed. We compare state-level benchmarking methods, when the benchmarking adjustment is (is not) part of the model. Supporting the mission of NASS, of providing timely, accurate, and useful statistics in service to US agriculture, the proposed method provides reliable, reproducible tools, including uncertainty measures for the point estimates.

In Section 2 we summarize the data available for this study, including the auxiliary acreage data sources, the remote sensing data and the weather data. In Section 3 we present the case study and data summaries. In Section 4 we introduce the proposed model, illustrated for a selected year-state-commodity combination and benchmarking methods. In Section 5 we present model-based estimation results.

2. Auxiliary Sources of Information

FSA is a USDA agency that administers U.S. Farm Programs authorized by the "Farm Bill" (USDA FSA 2014). For this, FSA collects data from farmers participating in such programs. The FSA county-level administrative data have been used by NASS ASB in constructing the official county-level estimates. The FSA data, of interest for our study, are the self-reported planted acreage values, aggregated at the county level.

RMA is a USDA agency that provides crop insurance to farmers participating in RMA's programs (USDA RMA 2014). The RMA county-level administrative data have also been used by NASS ASB in constructing the official county-level estimates. The RMA data, of interest for our study, are the self-reported failed acreage values, aggregated at the county level.

The Cropland Data Layer (CDL) is a NASS product that provides crop-specific land classification at 30 meter by 30 meter pixel resolution, covering the continental United States (USDA NASS 2016b). Remote sensing county-level data have also been used by NASS ASB in constructing the official county-level estimates. The remote sensing data, of interest for our study, are the CDL pixel counts, classified by commodity, aggregated at the county level.

We explore additional sources of auxiliary information, which are not currently being used by the NASS ASB in constructing the official county-level estimates. It is known that weather may determine the crop condition, from planting and harvesting dates, to critical stages in the crop growth impacting production. The weather data, of interest to our study, are variables from the National Oceanic and Atmospheric Administration (NOAA), aggregated at the district level. Finally, we explore the list frame control data available in the internal database, the NASS Enhanced List Maintenance Operations (ELMO).

3. Case Study

Although the ultimate goal is to provide reliable county-level estimates for all the state-commodity combinations in the U.S., in this study we present corn acreage estimates for Illinois, one of the largest production representative states in the Corn Belt and CAPS pilot program state, that has 102 counties and 9 ASDs. The case study state-commodity-year combination is Illinois-corn-2014.

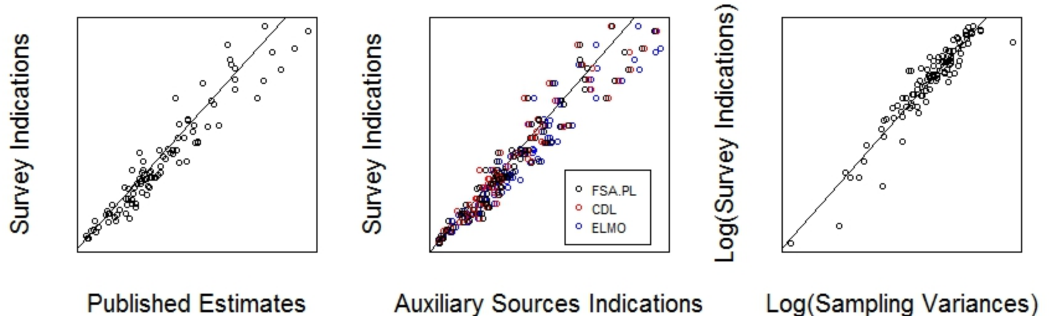
The state-level harvested corn acreage estimate is published in the NASS annual summary report, prior to the completion of data collection for CAPS, and serves as the benchmarking target for the county-level estimates to be published. The county-level harvested acreage survey indications and their estimated variances are provided by the CAPS summary. The acreage sources of auxiliary information are the FSA county-level planted corn acreage indications, the RMA county-level failed corn acreage indications, the CDL county-level acreage classified as corn, and the ELMO control corn acreage. All sources are available for the 102 counties in Illinois. The published county-level harvested corn acreage estimates are available in NASS QuickStats, at USDA NASS (2016c).

Harvested acreage depends on the planting date for the specific commodity. The planting date for corn varies from year to year, with optimum dates being the third and fourth weeks in April for most of the counties in Illinois. Some factors that determine the planting date for corn are the soil temperature and the soil moisture. After analyzing the NOAA weather data available, we decided to use the NOAA March Standardized Precipitation Index (SPI) as the weather source of auxiliary information. The NOAA March SPI is available for the 9 ASDs in Illinois.

3.1 Data Summaries

The FSA administrative data, the CDL remote sensing data, and the ELMO list frame data, considered in this study as covariates, are measures of planted acreage for the same county, within the state-commodity-year of interest. Moreover, it is known that one of the data sources used to produce the CDL is the FSA administrative acreage data, see Boryan et al. (2011). To avoid multicollinearity problems, we decided to include only one set of auxiliary acreage indications, as covariate observations, in the model. As a multicollinearity diagnostic, we considered the variance inflation factors (VIFs) and, for this study, the VIFs

Figure 1: County-level harvested acreage survey indications and county-level planted acreage auxiliary indications



From left to right, the first plot illustrates the strong linear relationship between the survey indications and the published estimates, the second plot illustrates the strong linear relationship between the survey indications and the auxiliary sources indications, and the third plot illustrates the strong relationship between the survey indications and the sampling variances, on the logarithmic scale. In the first two plots, we added the 45 *degrees line and in the third plot we added the best fitted line.*

for the different auxiliary sources are greater than 40. Also, the correlations between the survey indications and each of the auxiliary acreage data are greater than 0.95, suggesting the great efficiency of the selected auxiliary variables, with respect to explaining the variability in the harvested acreage variable.

The first two plots in Figure 1 illustrate the strong linear relationships between the survey indications of harvested acreage, the published estimates of harvested acreage and the multiple auxiliary sources indications of planted acreage. Due to under-coverage in CAPS, most of the county-level harvested acreage survey indications are lower than the corresponding published county-level estimates.

The county sample size is denoted by the number of records used to construct the county-level survey indications. The county population size is denoted by the number of records on the NASS list frame control data, ELMO. The county sample sizes range from 2 to 92, representing approximately 0.07% to 0.43% of the population sizes. The estimated CVs for the survey indications range from 9.9% to 92.3%, increasing with a decrease in the county sample size.

As mentioned in Section 1, the CAPS summary data include county-level estimates for the sampling variances, computed based on a delete-a-group Jackknife method, with 15 replicate weights. The county-level sampling variance increases with the decrease in the county sample size, and with the increase in the county-level harvested acreage, see first two plots in Figure 1. Moreover, we noticed a high correlation between the survey indications and the sampling variances, on the logarithmic scale, as illustrated in the last plot in Figure 1.

4. Models

The proposed model is a subarea-level model, where the area represents the ASD and the subarea represents the county. Of interest is estimation at the county and district level, while agreement between the county-level, district-level and state-level values is necessary. We acknowledge that the first, and most common, choice for modeling would be a Fay-Herriot model. However, it is known that the Fay-Herriot model does not account for the hierarchical structure of the data and the estimates constructed based on the Fay-Herriot model do not satisfy benchmarking constraints at higher levels. We consider a subarea-level model, that provides automatic agreement between the county-level and the district-level estimates. Also, in the Fay-Herriot model, the sampling variances are considered to be known. However, often, the sampling variances are estimated and incorporating the estimation error in the model is not straightforward. We extend the subarea-level model by adding a hierarchical level to model the sampling variances.

Let $i = 1, ..., m$ be an index for the m districts in the state, $j = 1, ..., n_{ci}$ be an index for the n_{ci} counties in the district i and n_{ij} be the county sample size. The total number of counties in the state is $\sum_{i=1}^{m} n_{ci} = n_c$ and the state sample size is $\sum_{i=1}^{m}\sum_{j=1}^{n_{ci}} n_{ij} = n$. The survey indication of harvested acreage in county i and district j is denoted by $\hat{\theta}_{ij}$, and its sampling variance is σ^2_{ij}. The county-level auxiliary information is $\mathbf{x}_{ij}$ and the district-level auxiliary information is $\mathbf{z}_i$. Let the hierarchical Bayesian subarea-level model be

$$\begin{aligned} n_{ij}^{-1}\hat{\theta}_{ij}|(\theta_{ij}, \sigma^2_{ij}) &\sim N(\theta_{ij}, n_{ij}^{-2}\sigma^2_{ij}) \\ \theta_{ij}|(\boldsymbol{\beta}, \sigma^2_u) &\sim N(\mathbf{x}_{ij}^{'}\boldsymbol{\beta}_1 + \mathbf{z}_i^{'}\boldsymbol{\beta}_2 + v_i, \sigma^2_u) \\ v_i|\sigma^2_v &\sim N(0, \sigma^2_v), \end{aligned} \tag{1}$$

where the county-level random effects are assumed to be independent, normally distributed with mean 0 and variance σ^2_u, and the district-level random effects are assumed to be independent, normally distributed with mean 0 and variance σ^2_v. Note that model (1) with known σ^2_{ij} and without the district-level effects, v_i, reduces to the area-level model, introduced by Fay and Herriot (1979).

It is important to note that the distributions of the county-level total acreage indications are skewed. In order to maintain the normality assumptions in (1), we let the observations on response variable be the survey indications of harvested acreage per unit, by dividing the county-level total acreage indication by the county-level sample size. As a consequence, the skewness is reduced and the symmetry of the distribution is improved.

In model (1), the sampling variances σ^2_{ij} are fixed. Often, estimates $\hat{\sigma}^2_{ij}$ are available, or can be computed, from the survey. Given that $\hat{\sigma}^2_{ij}$ are available, we further consider a more robust specification for the model (1), by adding an additional hierarchical level to the model, level corresponding to random sampling variances,

$$\begin{aligned} (n_{ij} - 1)\frac{\hat{\sigma}^2_{ij}}{\sigma^2_{ij}}|\sigma^2_{ij} &\sim \chi^2_{(n_{ij}-1)} \\ log(n_{ij}^{-2}\sigma^2_{ij})|(\boldsymbol{\alpha}, \sigma^2) &\sim N(log(\mathbf{x}_{ij})'\boldsymbol{\alpha}, \sigma^2) \end{aligned} \tag{2}$$

Wang and Fuller (2003), You and Chapman (2006), Gonzalez-Manteiga, et al. (2010), Erciulescu and Berg (2014) consider the case when the sampling variances are unknown and modeled separately. Note that the logarithmic transformation can be applied to the auxiliary acreage indications because they are strictly positive.

To complete the Bayesian model specification, we consider a priori independent parameters and noninformative, proper priors for $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma_u^2, \sigma_v^2, \sigma^2)$. Details on the specific form of the prior distributions are given in Section 5.

The Bayesian model (1-2) has numerous advantages. First, different sources of auxiliary information can be incorporated in a model-based approach to estimate crop acreage. Second, estimates can be easily obtained for any known function of the model parameters, for example parameter transformations due to benchmarking constraints, as well as estimates at other hierarchical levels of interest, for example district-level estimates. Third, the model accounts for the estimation error in the sampling variances. Fourth, posterior summaries, including credible intervals are automatically available.

4.1 Benchmarking

Let $n_{ij}\tilde{\theta}_{ij}$ be the hierarchical Bayesian estimate of the county-level acreage, for county j, district i. It is desirable that the sum of the county-level estimates within a district agrees with the district-level estimate, and that the sum of the district-level estimates in a state agrees with the state-level estimate. That is, if a denotes the state-level estimate and $\tilde{\theta}_{ij}^{Benchmarked}$ denotes the county-level estimate satisfying the constraint, the following relationship holds

$$a = \sum_{i=1}^{m} \sum_{j=1}^{n_{ci}} n_{ij} \tilde{\theta}_{ij}^{Benchmarked}.$$

Often, a is the sum of the area-level survey indications, and the benchmarking methods are known as internal benchmarking. However, a reliable state-level estimate, a may be available from a larger survey or from a previously published source, and preferred as the benchmarking constraint. In this study we consider external benchmarking methods, when a is known and fixed, a being the NASS previously published state-level estimate.

Widely used simple adjustments to the area-level estimates are available using the difference benchmarking and the ratio benchmarking. See Rao and Molina (2015) for an illustration of the two methods. Under the two methods, the area-level estimates are adjusted by a common factor, that does not depend on the area-specific precision. We apply both methods to the county-level estimates $n_{ij}\tilde{\theta}_{ij}$ and we also introduce an area-dependent difference benchmarking method, where the adjustment factor is inversely proportional to the area sample size. The idea behind this method is to assign a greater degree of reliability to the county-level estimates for large areas and greater adjustment to the small areas. The benchmarking adjustments are applied to the posterior distribution iterations of $\tilde{\theta}_{ij}$, and are not part of the initial model fit for the difference benchmarking and the ratio benchmarking.

Difference (DB)

$$a = \sum_{i=1}^{m} \sum_{j=1}^{n_{ci}} n_{ij} \tilde{\theta}_{ij}^{DB} \; ; \tilde{\theta}_{ij}^{DB} = \tilde{\theta}_{ij} + n^{-1} \left(a - \sum_{k=1}^{m} \sum_{l=1}^{n_{ck}} n_{kl} \tilde{\theta}_{kl} \right)$$

Alternative Difference (ADB)

$$a = \sum_{i=1}^{m}\sum_{j=1}^{n_{ci}} n_{ij}\tilde{\theta}_{ij}^{ADB} \,;\, \tilde{\theta}_{ij}^{DB} = \tilde{\theta}_{ij} + mn_{ij}{}^{-1}\left(a - \sum_{k=1}^{m}\sum_{l=1}^{n_{ck}} n_{kl}\tilde{\theta}_{kl}\right)$$

Ratio (RB)

$$\tilde{\theta}_{ij}^{RB} = \tilde{\theta}_{ij} \times a\left(\sum_{k=1}^{m}\sum_{l=1}^{n_{ck}} n_{kl}\tilde{\theta}_{kl}\right)^{-1}$$

The fourth benchmarking method considered is a parametric method. A parameter transformation is applied to the set of county-level parameters θ_{kl} before fitting the model. For more details on the Bayesian benchmarking (BB) method, see Nandram and Sayit (2011).

Parametric Transformation (BB)

$$(\theta_{11}, ..., \theta_{mn_{cm}}) \rightarrow (\theta_{11}, ..., \theta_{m(n_{cm}-1)}, \phi)\,,\, \phi = a - \sum_{k=1}^{m}\sum_{l=1}^{n_{ck}} n_{kl}\tilde{\theta}_{kl}$$

The *BB* method preserves the form of the joint normal distribution assumed for the original model parameters θ_{ij}. To sketch the derivation of the joint distribution of the transformed model parameters $(\theta_{11}, ..., \theta_{m(n_{cm}-1)}, \phi)$, let the subscript (mn_m) denote the deletion of the $(mn_m)^{th}$ element in a vector, let I_{n_c-1} denote the identity matrix of size $(n_c - 1)$, and let $\mathbf{n}$ denote the vector of subarea sample sizes. Then, under the constraint

$$a = \sum_{i=1}^{m}\sum_{j=1}^{n_{ci}} n_{ij}\theta_{ij},$$

the joint density for the first $(\sum_{i=1}^{m} n_{ci}) - 1$ subarea parameters is

$$\boldsymbol{\theta}_{(mn_{cm})} \sim MVN(\boldsymbol{\mu}_{(mn_{cm})}, \boldsymbol{\Sigma}_{(mn_{cm})}),$$

where

$$\begin{aligned}
\boldsymbol{\mu}_{(mn_{cm})} &= (\mathbf{x}, \mathbf{z})'_{(mn_{cm})}\boldsymbol{\beta} + \mathbf{v}_{(mn_{cm})} \\
&\quad + \mathbf{n}'_{(mn_{cm})}(\textstyle\sum_{i=1}^{m}\sum_{j=1}^{n_{ci}} n_{ij}^2)^{-1}\left(\textstyle\sum_{i=1}^{m}\sum_{j=1}^{n_{ci}} n_{ij}((\mathbf{x}_i, \mathbf{z}_i)'\boldsymbol{\beta} + v_i)\right) \\
\boldsymbol{\Sigma}_{(mn_{cm})} &= \sigma_u^2\left(I_{n_c-1} - (\textstyle\sum_{i=1}^{m}\sum_{j=1}^{n_{ci}} n_{ij}^2)^{-1}\mathbf{n}'_{(mn_{cm})}\mathbf{n}_{(mn_{cm})}\right).
\end{aligned}$$

Notice that this method is implemented using a nonsingular transformation (Vespers 2013) where the joint distribution of the first $(n_c - 1)$ subareas is modified, while the last subarea parameter is set to the difference between the state-level estimate and the sum of the $(n_c - 1)$ area means,

$$\theta_{mn_{cm}} = n_{mn_{cm}}^{-1} \left(a - \sum_{i=1}^{m} \sum_{j=1, j \neq n_{cm}}^{n_{ci}} n_{ij}\theta_{ij} \right).$$

5. Results

We fit models (1) and (1,2) for CAPS survey indications of harvested acreage per unit, separately. Each model is fit using no auxiliary information, and using each of the three auxiliary county-level acreage indications, FSA, CDL and ELMO, scaled by the county-level sample size, as covariate. The district-level weather information, NOAA SPI for March, is also incorporated in the models as covariate. Additional explanatory variables are used, defined as the difference between the planted acreage indications and the failed acreage indications, FSA-RMA, ELMO-RMA and CDL-RMA, and considered as covariates.

The prior distributions for the model parameters $(\boldsymbol{\beta}, \boldsymbol{\alpha})$ are Normal distributions with mean and variance denoted by the least squares estimates of $(\boldsymbol{\beta}, \boldsymbol{\alpha})$. The least squares estimates of $\boldsymbol{\beta}$ are obtained from fitting a simple linear model for the county-level survey indications against the county-level auxiliary indications. The least squares estimates of $\boldsymbol{\alpha}$ are obtained from fitting a simple linear model for the county-level sampling variances against the county-level auxiliary indications, on the logarithmic scale. The prior distributions for the model variance components σ_u^2, σ_v^2, and σ^2 are $Uniform(0, 10^8), Uniform(0, 10^8)$, and $Inverse - Gamma(10^3, 10^3)$, respectively.

The models are fit using R JAGS, and the posterior distributions constructed using Markov chain Monte Carlo simulation. We use 10000 Monte Carlo samples and 1000 burn-in samples, 3 chains, each thinned every 15 samples. The convergence is monitored using the multiple potential scale reduction factors. Also, once the simulated chains have mixed, we construct the effective number of independent simulation draws to monitor simulation accuracy.

5.1 Model-based Estimation Results

We denote the model-based estimate (ME) of θ_{ij}, under no benchmarking constraints, by $\tilde{\theta}_{ij}$, and compute it as the posterior mean of θ_{ij}. The state-level simple benchmarking adjustments, when the adjustment is not part of the model fit, are applied to the value of the parameter θ_{ij} at each iteration, constructing the simulated chains of $\theta_{ij}^{MEDB,MEADB,MERB}$, under the DB, ADB, RB benchmarking constraints, respectively. The parametric benchmarking results in simulated chains of θ_{ij}^{MEBB}. The model-based estimates (MEDB, MEADB, MERB, MEBB) of θ_{ij}, under the (DB, ADB, RB, BB) benchmarking constraints, are denoted by $\tilde{\theta}_{ij}^{MEDB,MEADB,MERB,MEBB}$, and computed as the posterior mean of $\theta_{ij}^{MEDB,MEADB,MERB,MEBB}$. The i^{th} district-level estimate is the posterior mean of the iterations on aggregated $\left(\theta_{ij}, \theta_{ij}^{MEDB,MEADB,MERB,MEBB}\right)$ iterations within district i, under no benchmarking adjustment, and under the different benchmarking adjustments, respectively. Similarly, estimated variances of the model-based estimates are constructed as the posterior variances of the corresponding parameters.

We compare the harvested acreage county-level model-based estimates, under the different benchmarking methods to the survey indications and to the published estimates.

Metrics to quantify the relative differences between the different point estimates are given in the form of

$$PM_{Source} := \frac{\tilde{\theta} - Source}{Source},$$

where *Source* refers to *CAPS*, for the survey indications, and to *Published*, for the published estimates. Also, the relative difference between the estimated variances of the survey indications and the estimated variances of the model-based estimates, is quantified by

$$VM_{CAPS} := \frac{v\hat{a}r(\tilde{\theta}) - v\hat{a}r(\hat{\theta})}{v\hat{a}r(\hat{\theta})}.$$

Numerical results for county-level estimates are presented in Tables 1, 3 and 5, and for district-level estimates are presented in Tables 2, 4 and 6. The last set of three rows in Table 1, 2, 3, 4, 5 and 6, correspond to summaries of the estimated coefficients of variation (CVs) for the different point estimates constructed.

Using the results in Tables 1 and 2, we compare the model performance under different sources of auxiliary information, to the model performance under no auxiliary information, to illustrate the degree of selection of auxiliary data. The model-based estimates, based on the subarea-level model with no benchmarking constraint and no covariate information are closer to the survey indications, than the model-based estimates with no benchmarking constraint and FSA, ELMO or CDL as covariate information. There is a great reduction in the estimated variance and in the estimated CV for the model-based estimates, when auxiliary information is considered versus when no auxiliary information is considered.

The model-based estimates, under no benchmarking constraint, are mostly lower than the survey indications and than the published estimates, results being consistent across the models using the three different auxiliary sources. The model-based estimates, under the benchmarking constraints, are within less than 1.4% of the published estimates. The greatest difference between the model-based estimates and the published estimates is for the alternative difference adjustment (ADB), that is sensitive to the county sample size. The median absolute relative difference, to the published estimates, in the model-based estimates is larger for the estimates constructed under no benchmarking constraints, than for the estimates constructed under the different benchmarking methods, with smaller median absolute relative difference for the ratio benchmarking. The median absolute relative difference in the estimated variance of the model-based estimates is similar for all the benchmarking methods, and above 70%.

The minimum, median and maximum values of the estimated CVs of the county-level survey indications are $9.9, 19.2,$ and 92.3, respectively. The minimum, median and maximum values of the estimated CVs of the district-level survey indications are $4.5, 6.7,$ and 8.7, respectively. The model-based estimates, using auxiliary information, have medians of estimated CVs approximately three times lower than the CVs of the survey indications. Under the difference benchmarking methods, the largest estimated CVs of the model-based county-level estimates are approximately six times smaller than the largest estimated CVs for the county-level survey indications, and the largest estimated CVs of the model-based district-level estimates are approximately two times smaller than the largest estimated CVs for the district-level survey indications. The parametric benchmarking method leads to estimated CVs that are larger than the corresponding values for the simple benchmarking methods, when the adjustment is not part of the model fit, but still lower than the CVs of the survey indications.

Table 1: 2014 Illinois Corn Harvested Acreage: Properties of the Estimates (%)
County-level Survey Indications and Model-Based Estimates
Different Sources of Auxiliary Information; No Benchmarking

Metric	Statistic	no covars	FSA	ELMO	CDL	FSA-RMA	ELMO-RMA	CDL-RMA
PM$_{CAPS}$	min	-42.2	-74.3	-42.6	-69.3	-74.3	-42.3	-69.4
	median	0.2	-4.0	-2.2	-3.8	-4.1	-2.9	-3.6
	max	52.6	80.0	47.5	55.8	80.6	47.4	56.5
PM$_{Published}$	min	-42.9	-24.4	-34.3	-25.7	-25.1	-34.4	-26.2
	median	-11.5	-13.4	-12.9	-13.6	-13.4	-12.9	-13.6
	max	65.2	10.1	22.4	13.8	8.5	22.0	15.1
VM$_{CAPS}$	min	-93.0	-99.7	-99.8	-99.6	-99.7	-99.8	-99.6
	median	-20.7	-85.5	-90.7	-83.6	-85.5	-90.8	-84.3
	max	-0.6	-18.9	-28.4	-9.0	-19.8	-32.4	-14.6
CV	min	9.9	4.3	3.7	4.5	4.4	3.6	4.4
	median	17.6	7.5	6.2	7.9	7.5	5.9	7.9
	max	50.0	31.7	27.3	31.2	31.9	26.3	30.5

Table 2: 2014 Illinois Corn Harvested Acreage: Properties of the Estimates (%)
District-level Survey Indications and Model-Based Estimates
Different Sources of Auxiliary Information; No Benchmarking

Metric	Statistic	no covars	FSA	ELMO	CDL	FSA-RMA	ELMO-RMA	CDL-RMA
PM$_{CAPS}$	min	-11.7	-15.2	-7.1	-12.7	-15.3	-7.3	-12.8
	median	-6.2	-5.0	-4.8	-5.4	-4.8	-4.8	-5.4
	max	-0.8	-2.9	-2.4	-3.5	-3.0	-2.3	-3.5
PM$_{Published}$	min	-20.3	-14.8	-16.6	-16.3	-14.6	-16.4	-16.2
	median	-14.8	-13.6	-13.7	-13.8	-13.7	-13.8	-13.7
	max	-8.9	-12.3	-4.2	-11.0	-12.3	-4.2	-11.2
VM$_{CAPS}$	min	-47.8	-85.0	-83.5	-80.2	-83.3	-84.5	-80.5
	median	-23.6	-73.8	-70.9	-68.2	-72.9	-72.7	-69.7
	max	47.8	-50.7	-44.0	-39.4	-46.5	-48.3	-45.4
CV	min	4.8	3.3	3.3	3.4	3.2	3.1	3.4
	median	6.3	3.5	3.6	3.8	3.5	3.5	3.7
	max	7.7	5.0	4.8	5.6	4.9	5.0	5.4

Table 3: 2014 Illinois Corn Harvested Acreage: Properties of the Estimates (%)
County-level Survey Indications and Model-Based Estimates
Auxiliary Information: FSA and NOAA March SPI; Different Benchmarking Methods

Metric	Statistic	ME	MEDB	MEADB	MERB	MEBB
PM$_{CAPS}$	min	-73.8	-62.4	-33.0	-69.6	-77.0
	median	-4.1	14.0	16.3	11.1	9.8
	max	88.5	239.7	1204.7	118.5	70.1
PM$_{Published}$	min	-28.3	-9.5	-11.1	-16.9	-35.0
	median	-13.6	1.2	1.1	0.1	-1.4
	max	5.9	110.4	228.6	22.8	20.7
VM$_{CAPS}$	min	-99.7	-99.7	-99.4	-99.7	-99.1
	median	-87.8	-88.8	-88.4	-85.2	-70.4
	max	66.5	96.1	1185.9	125.8	131.8
CV	min	4.1	3.3	3.4	3.4	4.7
	median	6.9	5.6	5.8	6.7	9.6
	max	32.2	18.1	13.5	32.2	56.8

Table 4: 2014 Illinois Corn Harvested Acreage: Properties of the Estimates (%)
District-level Survey Indications and Model-Based Estimates
Auxiliary Information: FSA and NOAA March SPI; Different Benchmarking Methods

Metric	Statistic	ME	MEDB	MEADB	MERB	MEBB
PM_{CAPS}	min	-15.3	-1.8	1.1	-1.9	2.4
	median	-5.0	9.8	8.9	10.1	9.5
	max	-1.6	29.4	35.7	14.1	11.8
$\text{PM}_{Published}$	min	-15.3	-7.0	-7.9	-1.8	-6.2
	median	-13.3	0.1	-0.5	0.5	-0.2
	max	-11.1	16.9	22.5	3.1	4.3
VM_{CAPS}	min	-84.7	-89.7	-89.7	-86.2	-81.7
	median	-71.0	-80.1	-79.9	-75.8	-69.4
	max	-46.9	-42.7	-33.6	-39.5	-38.4
CV	min	3.2	2.1	2.2	2.4	2.6
	median	3.6	2.7	2.7	3.1	3.0
	max	5.2	4.4	4.5	5.1	5.4

Table 5: 2014 Illinois Corn Harvested Acreage: Properties of the Estimates (%)
County-level Survey Indications and Model-Based Estimates
Auxiliary Information: ELMO/CDL and NOAA March SPI; Different Benchmarking Methods

			with	ELMO				with	CDL		
Metric	Statistic	ME	MEDB	MEADB	MERB	MEBB	ME	MEDB	MEADB	MERB	MEBB
PM_{CAPS}	min	-42.4	-31.6	-30.0	-33.8	-39.2	-69.3	-58.1	-35.5	-64.5	-72.7
	median	-1.9	14.7	15.1	12.7	11.3	-3.1	15.6	15.8	12.0	10.2
	max	45.6	191.1	1099.4	67.2	57.1	75.6	233.8	1182.8	103.1	61.1
$\text{PM}_{Published}$	min	-15.8	-7.2	-8.1	-3.3	-8.5	-16.6	-6.9	-7.8	-3.6	-7.4
	median	-13.7	-0.1	-0.7	-0.8	-0.6	-13.4	0.3	-1.5	0.2	-1.2
	max	-4.3	14.5	19.9	9.9	12.5	-11.2	14.2	19.8	2.7	7.0
VM_{CAPS}	min	-99.8	-99.8	-99.5	-99.8	-99.4	-99.7	-99.7	-99.3	-99.6	-99.3
	median	-90.9	-91.7	-91.5	-89.4	-79.7	-87.4	-87.8	-87.5	-84.2	-75.5
	max	-14.0	15.2	1133.1	15.0	31.4	59.7	84.9	1204.7	114.8	46.5
CV	min	3.7	3.1	3.0	2.8	3.9	4.2	3.4	3.5	3.4	4.2
	median	6.1	5.0	5.0	5.8	8.0	7.2	6.0	6.0	7.0	8.8
	max	29.2	16.2	12.9	29.4	47.9	31.7	17.9	13.5	31.7	44.3

Table 6: 2014 Illinois Corn Harvested Acreage: Properties of the Estimates (%)
District-level Survey Indications and Model-Based Estimates
Auxiliary Information: ELMO/CDL and NOAA March SPI; Different Benchmarking Methods

			with	ELMO				with	CDL		
Metric	Statistic	ME	MEDB	MEADB	MERB	MEBB	ME	MEDB	MEADB	MERB	MEBB
PM_{CAPS}	min	-7.2	3.1	2.1	6.6	1.7	-13.1	0.3	2.7	0.5	2.9
	median	-5.0	9.0	9.5	9.1	9.5	-4.5	9.9	8.0	10.4	8.4
	max	-2.5	26.7	32.6	12.0	11.7	-3.2	26.3	32.5	11.9	12.9
$\text{PM}_{Published}$	min	-14.4	-7.2	-8.0	-1.8	-3.2	-16.7	-6.6	-7.5	-3.7	-3.3
	median	-13.5	-0.5	-1.1	-0.7	-0.2	-13.3	0.2	-1.2	0.3	-0.3
	max	4.3	26.4	39.6	19.8	22.5	6.8	30.3	44.1	23.5	16.5
VM_{CAPS}	min	-82.2	-89.0	-89.3	-85.8	-85.4	-78.5	-86.6	-85.5	-82.3	-82.7
	median	-68.0	-78.8	-77.7	-75.0	-76.3	-67.1	-77.5	-74.6	-70.0	-72.6
	max	-45.1	-40.0	-29.4	-42.2	-46.2	-39.7	-39.8	-30.1	-37.6	-44.5
CV	min	3.2	2.2	2.2	2.5	2.4	3.4	2.3	2.5	2.6	2.5
	median	3.7	2.7	2.7	2.9	2.9	3.8	2.9	2.9	3.4	2.9
	max	5.0	4.6	4.7	5.1	5.1	5.4	4.6	4.7	5.3	5.2

5.2 Model Comparison

For comparison, we fit a total of 50 models for county-level harvested acreage, 26 of the form (1), corresponding to fixed sampling variances, and 24 of the form (1,2), corresponding to random sampling variances. For each case, fixed and random sampling variances, the models are fit using the initial parameterization and using the BB parameterization, and incorporating different sets of covariates illustrated in the first two columns of Table 7. That is, model (1) is fit using no auxiliary data, using either source of acreage indications described in Section 2 and in the first paragraph in Section 5 (FSA, ELMO, CDL, FSA-RMA, ELMO-RMA, or CDL-RMA), and using either source of acreage indications in combination with the weather data source (NOAA). Similarly, model (1,2) is fit using either source of acreage indications, alone, and in combination with the weather data; we do not fit model (1,2) with no covariates.

Our choice of model comparison is the Deviance Information Criterion (DIC). While lower DIC indicates better fit, the model comparison using DIC depends on the specific formulation of the distribution, hence it is not applicable when different parameterizations exist. For this reason, we compare the contribution of the different sources of auxiliary information for each case, fixed and random sampling variance, under each parameterization, initial and BB. See Table 7 for the DICs for all the 50 fitted models. Columns three and four of Table 7 illustrate results for model (1) with different sets of auxiliary information, under the two parameterizations, initial and BB, respectively. Similarly, columns five and six of Table 7 illustrate results for model (1,2), with different sets of auxiliary information, under the two parameterizations, initial and BB, respectively.

The goodness of fit for the harvested acreage survey indications increases when auxiliary information is incorporated in the model, the best fit being when the ELMO data is used as a covariate. The reason we presented results, in previous sections, based on the model (1,2) with FSA as covariate, was to provide a fair comparison to the current NASS method of constructing county-level estimates; as described in Section 2, the current method uses the FSA data, and it does not use the ELMO data.

Table 7: 2014 Illinois Corn Harvested Acreage: DIC Subarea-Level Models Incorporating Auxiliary Sources of Information

Auxiliary Information		Model (1)	Model (1), BB	Model (1,2)	Model (1,2), BB
County-level	District-level				
-	-	1592.3	1604.4	-	-
FSA	-	1517.4	1582.3	3998.0	4074.9
ELMO	-	1480.1	1560.8	3956.3	4042.9
CDL	-	1521.3	1575.8	3995.8	4062.0
FSA-RMA	-	1519.5	1580.9	3993.8	4072.9
ELMO-RMA	-	1477.9	1552.9	3955.4	4044.2
CDL-RMA	-	1516.2	1582.3	3997.8	4058.5
FSA	NOAA	1516.0	1576.0	3993.3	4070.0
ELMO	NOAA	1481.1	1552.5	3957.8	4039.4
CDL	NOAA	1513.2	1572.7	3988.8	4056.2
FSA-RMA	NOAA	1516.5	1577.1	3982.2	4066.5
ELMO-RMA	NOAA	1478.6	1557.1	3959.7	4045.2
CDL-RMA	NOAA	1513.3	1577.0	3988.3	4070.6

Models incorporating the FSA data or the CDL data result in similar performance. The model goodness of fit increases slightly when the RMA failed acreage indication is subtracted from the planted acreage indication (either FSA, ELMO or CDL), in comparison to the model fit incorporating the planted acreage indication alone. Also, there is a slight increase in the goodness of fit for the model with two covariates, an acreage indication and the weather data, in comparison to the model incorporating the acreage indication alone.

6. Conclusions

We explore a range of auxiliary information available at different levels, county and district, that provides a good source of explanatory data for model-based estimation of county-level harvested acreage survey indications. Preliminary spatial analysis, not included in this paper, indicates that the presence of auxiliary information as covariate information in the model reduces the spatial effect noticed in the survey county-level indications.

We propose a novel subarea-level model to construct reliable county-level harvested acreage estimates, having CVs approximately 65% lower than the CVs of the survey indications. The coefficients of variation for the county-level survey indications are reduced when borrowing strength from all counties within a district and all districts within a state. Further reduction is observed when auxiliary information is included in the models. Under a simple difference benchmarking method, the largest such coefficient of variation is less than 13.5%. While the survey indications are lower than the published estimates, additional analysis shows that the 95% credible intervals of the model-based estimates, under the benchmarking constraints, cover the published estimates.

The methods illustrated in this paper are a good fit for a selected state-year-commodity combination, when the parameters of interest are county-level and district-level harvested acreage estimates. The methods can be applied to any state-year-commodity combination, as long as the auxiliary data are available. Similar results are obtained when the auxiliary failed acreage indications are used to construct covariate information as the difference between planted acreage and failed acreage.

The proposed model performs well for improving planted acreage survey indications and results, not included in this paper, are similar to the results for the model-based harvested acreage estimates, presented in Section 5. Additional analysis on the comparison between the proposed model and the current method of estimating county-level acreage, shows that the contribution of the FSA indications to the county-level planted acreage estimate is similar for the two methods; approximately, a weight of 80% is assigned to the FSA indications and a weight of 20% is assigned to the survey indications. However, the additional benefit of the proposed model-based estimation is that the contribution of different auxiliary information can be evaluated, and that this contribution is county-specific, in contrast with the current method where an initial weight of 80% is allocated to the FSA indications, for all the counties, and for all the states.

Future work includes exploring additional sources of auxiliary information, and constructing model-based county-level estimates for all the sampled states and commodities, in a given year. Given any two of the three quantities, harvested acreage, production and yield, the third may be computed as a function of the two; yield is considered to be the ratio of production to harvested acreage. While the proposed model is a good fit for the yield survey indications and for the production survey indications, under a simple transformation, challenges arise when the auxiliary sources of information are evaluated and when nonlinear benchmarking constraints are imposed on the yield (ratio) estimates. Finally, the effect of restricting the parameter space in the BB benchmarking method leads to an artificial reduction in the variance of the parameter of interest. Also, for a set of heterogeneous subareas, *choosing* the last subarea to apply the transformation to the rest $n_c - 1$ subareas leads to different results in the posterior distributions of the subarea means. In future work we will explore alternative benchmarking methods.

REFERENCES

Adrian, D.W. (2012), "A model-based approach to forecasting corn and soybean yields", *Proceedings of the International Conference On Establishment Surveys*, 302190.

Battese G.E., Harter R.M., Fuller W.A. (1988), "An error component model for prediction of county crop areas using survey and satellite data," *Journal of the American Statistical Association*, 83, 28-36.

Bellow M.E. (1993), "Application of satellite data to crop area estimation at the county level," *USDA NASS Report*, June.

Bellow M.E. 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.

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

Bellow M.E. and Lahiri P. (2012), "Evaluation of Methods for County Level Estimation of Crop Harvested Area that Employ Mixed Models" *Proceedings of the DC-AAPOR / WSS Summer Conference, American Statistical Association*, Bethesda Maryland.

Boryan C., Yang Z., Muller R. and Craig M. (2011), "Monitoring US agriculture: the US Department of Agriculture, National Agricultural Statistics Service, Cropland Data Layer Program" *Geocarto International*, 1-18, iFirst article.

Erciulescu A.L. and Berg E. (2014), "Small Area Estimates for the Conservation Effects Assessment Project," *Frontiers of Hierarchical Modeling in Observational Studies, Complex Surveys and Big Data: A Conference Honoring Professor Malay Ghosh, College Park, MD; Women in Statistics Conference, Cary, NC*.

Fay R.E. and Herriot R.A. (1979), "Estimates of income for small places: an application of James-Stein procedures to census data," *Journal of the American Statistical Association*, 74, 269-277.

Fuller W.A. and Battese G.E. (1973), "Transformations for Estimation of Linear Models with Nested-Error Structure," *Journal of the American Statistical Association*, 68, 626-632.

Fuller W.A. and Goyeneche J.J. (1998), "Estimation of the state variance component," *Unpublished manuscript*.

Gonzalez-Manteiga W., Lombardia M.J., Molina I., Morales D. and SantaMaria L. (2010), "Small area estimation under Fay-Herriot Models with nonparametric estimation of heteroscedasticity," *Statistical Modelling*, 10, 215-239.

Kott P.S. (1998), "Using the delete-a-group Jackknife variance estimator in practice," *JSM Proceedings, Survey Research Methods Section. Alexandria, VA: American Statistical Association*, 763-768.

Nandram B. and Sayit H. (2011), "A Bayesian analysis of small area probabilities under a constraint," *Survey Methodology*, 37, 2, 137-152.

Rao J.N.K. and Molina I. (2015), "Small Area Estimation," *Wiley Series in Survey Methodology*.

Torabi M. and Rao J.N.K. (2014), "On small area estimation under a subarea level model," *Journal of Multivariate Analysis*, 127, 36-55.

US Census Bureau (2016), " Small Area Income and Poverty Estimates,"
https://www.census.gov/did/www/saipe/data/.

USDA FSA (2014), "Farm Bill Home," *http://www.fsa.usda.gov/programs-and-services/farm-bill/index*.

USDA NASS (2016a), "Publications: Agricultural Statistics, Annual,"
https://www.nass.usda.gov/Publications/Ag_Statistics.

USDA NASS (2016b), "CropScape and Cropland Data Layer,"
https://www.nass.usda.gov/Research_and_Science/Cropland/SARS1a.php.

USDA NASS (2016c), "QuickStats," *https://quickstats.nass.usda.gov/*.

USDA RMA (2014), "THE FARM BILL," *http://www.rma.usda.gov/news/currentissues/farmbill/*.

Vespers A.J. (2013), "Three Essays of Applied Bayesian Modeling: Financial Return Contagion, Benchmarking Small Area Estimates, and Time-Varying Dependence," *Doctoral dissertation, Harvard University*.

You Y. and Chapman B. (2006), "Small area estimation using area level models and estimated sampling variances," *Survey Methodology*, 32, 97-103.

Wang J. and Fuller W.A. (2003), "The mean squared error of small area predictors constructed with estimated area variances," *Journal of the American Statistical Association*, 98(463), 716-723.