Hierarchical Bayesian Model with Inequality Constraints for County Estimates

Lu Chen[*], Balgobin Nandram[†] and Nathan B. Cruze [‡]

Abstract

In the production of US agricultural official statistics, certain inequality and benchmarking constraints must be satisfied. For example, available administrative data provide an accurate lower bound for the county-level estimates of planted acres, produced by the US Department of Agriculture's (USDA) National Agricultural Statistics Services (NASS). In addition, the county-level estimates within a state need to add to the state-level estimates. A sub-area hierarchical Bayesian model with inequality constraints to produce county-level estimates that satisfy these important relationships is discussed, along with associated measures of uncertainty. This model combines the County Agricultural Production Survey (CAPS) data with administrative data. Inequality constraints add complexity to fitting the model and present a computational challenge to a full Bayesian approach, so improved performance is needed to justify the additional computational burden. To evaluate the inclusion of these constraints, the models with and without inequality constraints were compared using 2014 corn planted acres estimates for two states. The performance of the model with inequality constraints illustrates the improvement of county-level estimates in accuracy and precision while preserving required relationships.

Key Words: Administrative Data, Agricultural Statistics, Bayesian Diagnostic, Benchmarking, Small Area Estimation, Survey Data

1. Introduction

The National Agricultural Statistics Service (NASS), the primary statistical data collection agency within the U.S. Department of Agriculture (USDA), conducts the County Agricultural Production Survey (CAPS) annually. CAPS provides county-level estimates for crops by commodity: planted acres, harvested acres, yield and production. The current method of producing these official estimates is an expert assessment conducted by NASS's Agricultural Statistics Board (ASB), which incorporates multiple sources of information. The information includes the CAPS estimates and administrative data whenever it is available. These county-level estimates are key indicators to farmers, ranchers and a number of federal and state agencies for decision making. Two USDA agencies, the Farm Service Agency (FSA) and the Risk Management Agency (RMA), consider the estimates as part of their processes for distributing farm subsidies and insurance, respectively.

In the current process of setting official statistics, the ASB analyzes the survey estimates and integrates them with multiple data sources. To arrive at official estimates, the ASB relies on standard processes, multiple data sources, historical performance of these sources, and expert judgment. In a statistical sense, the ASB results are not reproducible and measures of uncertainty cannot be produced. Given the importance of the crops county estimates program, NASS engaged a panel of experts under the National Academies of Sciences, Engineering, and Medicine (NASEM) for guidance and recommendations on implementing small area models for integrating multiple sources of information to provide more precise county-level crop estimates with measures of uncertainty.

[*]National Institute of Statistical Sciences and USDA National Agricultural Statistics Service.

[†]Worcester Polytechnic Institute, Department of Mathematical Sciences and USDA National Agricultural Statistics Service.

[‡]USDA National Agricultural Statistics Service.

In recent years, small area models have gained increased attention by academic researchers and government agencies. The small area estimation models can "borrow" strength from related areas across space and/or time or through auxiliary information to provide "indirect" but reliable estimates for small areas while also increasing the precision. One challenge of producing a model-based approach is the ability to provide reliable and coherent estimates that satisfy important relationships nested among estimates and administrative data. The NASS county-level official estimates of planted acres should "cover" the corresponding available administrative data while also satisfying benchmarking constraints so that county-level estimates add up to the state-level estimates. In this paper, incorporating constraints of the planted acres estimates into small area models are applied and hierarchical Bayesian models with constraints for small area estimation are discussed. Before NASS can adopt a model-based approach to producing crops county estimates, the model must incorporate all known relationships. Publishing model-based estimates will lead to improved reproducibility and transparency.

Two major types of small area models, area-level and unit-level models, have been developed based on both frequentist and Bayesian methods. Pfeffermann (2013) and Rao and Malina (2015) provide a comprehensive overview of the development, methods and application of small area estimation including various types of area-level and unit-level models. For continuous responses, the first and most common model would be the Fay-Herriot model (Fay and Herriot, 1979) in small area estimation. It is an area-level model based on a "Normal-Normal-Linear" assumption. That is, the direct estimates and area-level random effects are both assumed to follow normal distribution and a linear regression function relates the true estimates of interest to covariates. The popular unit-level model, nested-error regression (NER) model, is proposed by Battese, Harter and Fuller (1988) when data are available on the individual sampled units. The NER model is also developed under the normality assumption.

Recent studies and papers related to the NASS crops county estimates program have shown that the hierarchical Bayesian small area models can incorporate auxiliary sources of data to improve county-level survey estimation of crop totals with measures of uncertainty. Battese, Harter and Fuller (1988) introduced the unit-level models for small area estimation based on nested error linear regression. They combined survey indications with satellite data. Erciulescu, Cruze and Nandram (2019) proposed and implemented a double shrinkage hierarchical Bayesian sub-area level model to provide the acreage estimates with associated measures of uncertainty. The paper discussed the results when integrating different data sources and showed that the county-level model-based acreage estimates decreased the coefficients of variance relative to the survey ones. Erciulescu, Cruze and Nandram (2020) discussed the challenges of missing data, either survey responses or administrative data, when fitting the hierarchical Bayesian sub-area level model to obtain the crops total estimates for the whole nation. In these two papers, the state-to-county benchmarking constraint is handled.

However, the inequality constraint problems have not been addressed in the aforementioned literature. Cruze et al. (2019) identified these constraints among estimates and administrative data as a necessity. Nandram, Cruze and Erciulescu (2020) addressed the inequality constraint problem and proposed several hierarchical Bayesian models for NASS crops county-level planted estimates. They discuss the methodologies of fitting constrained models and provide a simulation study to show the performance of all models.

In this paper, the models with inequality constraints by Nandram, Cruze and Erciulescu (2020) are applied to 2014 NASS CAPS data. The challenges for providing constrained estimates of planted acres in small area agricultural models are discussed. In Section 2, input data sources and some particular needs of the NASS crops county estimates for total

planted acres are presented. Section 3 presents the hierarchical Bayesian models with inequality constraints to produce reliable and coherent county-level estimates and associated measures of uncertainty. The external ratio benchmarking is applied to the county-level estimates so that they sum to state targets. The results are contrasted with those obtained from unconstrained models. In Section 4, a case study based on two different states shows the model-based estimation results and highlights the different performances of the constrained models and the unconstrained models. Conclusions and future work are presented in Section 5.

2. Data Sources and Requirements

2.1 Survey Data

Although NASS has been producing official county-level crop inventories since 1917, it was in 2011 that NASS completely implemented the large-scale probability survey, CAPS, to provide county-level official estimates for many principle small grains and row crops in several states. In 2012, CAPS was implemented in all eligible states. The list of crops and states in CAPS may change year to year depending on the requirement of coverage for federally mandated program crops and others. Figure 1 shows the 2019 CAPS states. The row crops (e.g. corn, soybeans) CAPS was conducted in 41 states shown in blue and orange. The small grains (e.g. barley, oats) CAPS was conducted in 32 states shown in blue and light yellow. All other states (shown in red) were not included in 2019 CAPS.

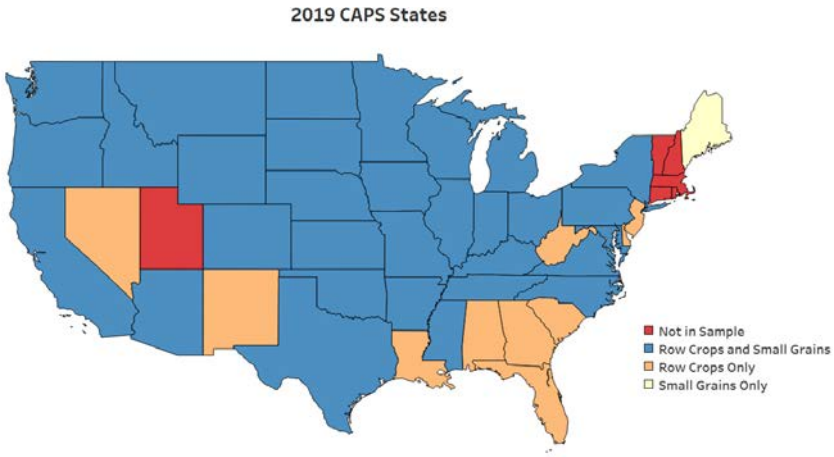

Figure 1: 2019 row crops and small grains CAPS states

As discussed in the introduction, the smallest area of CAPS is the county. Historically, NASS has also produced estimates for an intermediate domain called the agricultural statistics district (ASD). Each ASD is comprised of contiguous counties within the state. Both county-level and ASD level survey estimates and associated variance estimates are available in CAPS summary. The state-level estimates of planted acres are published before the completion of data collection for the CAPS. These estimates provide the benchmarking state targets for the county-level estimates to be published later.

2.2 Auxiliary Data

NASS obtains auxiliary sources of information on crop acres from FSA and RMA. Both agencies have farmer-reported administrative data on planted acres. FSA programs are popular but not compulsory. Farmers who participate in FSA programs certify the acres

planted by crop type. However, the participation rates vary by state. For example, for corn commodity, the participation rates are higher in the corn-belt states than some of the other states. Because the data are certified by farmers, the FSA data on planted acres serve as the lower bound of NASS published statistics of planted acres.

RMA receives administrative data on planted acres through independent crop insurance agents whenever farmers file claims that are associated with these programs. The participation rates of RMA vary by state and commodity. Farmers may not participate in the crop insurance program or they may not insure all crop commodities grown. Therefore, NASS treats the RMA administrative data on planted acres as a lower bound on the planted acreage within each county.

Because NASS treats both FSA and RMA data as the lower bounds of the county-level planted acreage estimates, the definition of the lower bound in the constrained models is the maximum of both sources of administrative data. That is, where FSA and RMA acreages may differ, the larger is taken as a lower bound, and the smaller is satisfied as a result.

2.3 Important relationships for planted acres

In the production of the official statistics for total acres reported by NASS, certain inequality and benchmarking constraints should be satisfied. NASS's official estimates of planted acres should "cover" corresponding available administrative data: FSA and RMA planted acreage data within any given geographic boundary, such as the US, a state, and a county. The relationship of NASS official statistics and FSA administrative data of total planted acreage for corn at US level from 2012 to 2019 is displayed in Figure 2 (reproduced from USDA, 2019). The final planted and failed acreage reported to FSA, final planted acreage from NASS, and the difference between FSA reported acreage and the NASS planted acreage for corn are displayed. The bars show that the differences between NASS official estimates and FSA data are all positive at the US level. NASS currently uses the top-down method to produce official county-level estimates that satisfy the county-state benchmarking constraint. However, the county-level survey estimates of the planted

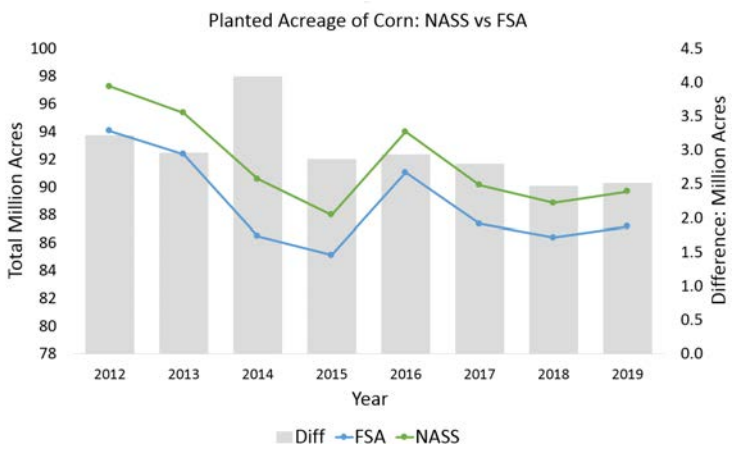

Figure 2: The US-level planted acreage estimates of corn for NASS and FSA

acreage do not always satisfy the constraints.

Figure 3 indicates that the points in plot of the survey estimates versus the FSA data are scattered around the 45 degree line. Some of the survey estimates are one or two standard deviations below the corresponding FSA or RMA data. This introduces difficulties for model estimates to preserve the relationships for the basic small area models without constraints. However, the inequality constraints must be incorporated into the model so

the all known relationships are satisfied at all levels before NASS can rely on model-based estimates as the foundation for the final official estimates.

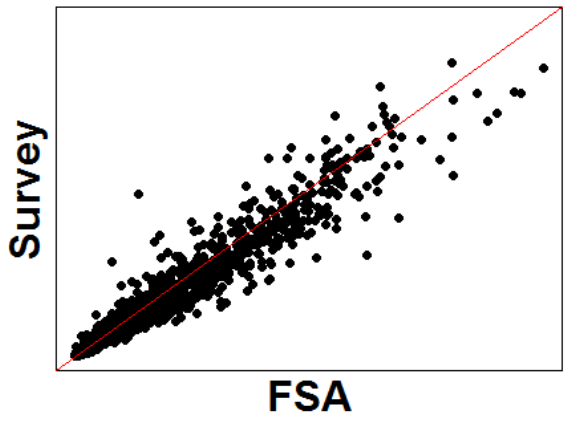

Figure 3: The county-level planted acreage estimates of 2014 corn for CAPS and FSA in all eligible counties

3. Models

Bayesian area-level and sub-area level models are popular in small area estimation. Models with constraints are considered based on the original small area models (Rao and Molina 2015; Erciulescu, Cruze and Nandram, 2020). In this section, models, with and without constraints, are presented and applied in a case study of 2014 corn data. The area-level model without inequality constraints is first introduced by Fay and Herriot (1979), where an area represents a county. The sub-area level models without inequality constraints are discussed by Fuller and Goyeneche (1998) and Torabi and Rao (2014) as an extension of FH model. In the sub-area level models, an area is an ASD and a subarea is a county. Nandram, Cruze and Erciulescu (2020) propose and discuss both area and sub-area level models to address the inequality problems into the models.

3.1 Models without Constraints

Erciulescu, Cruze and Nandram (2020) discuss and apply the hierarchical Bayesian sub-area model to estimate the number of planted and harvested acres. In their paper, the county-state benchmarking constraint is handled by ratio benchmarking in the output analysis but the inequality constraints are not addressed either in the model or in the output analysis. In this paper, this model is referred as the model without constraints and several comparisons between this type of model and models with constraints will be presented in Section 4.

Suppose that there are n counties in one state. Let $i = 1, \ldots, m$ be an index for m ASDs in the state and $j = 1, \ldots, n_i$ be an index for the county within the i^{th} ASD. The survey estimate of planted acreage in county i and district j is denoted by $\hat{\theta}_{ij}$ and the

associated survey variance is $\hat{\sigma}_{ij}^2$. The auxiliary data used in the models are $\boldsymbol{x}_{ij}$, including an intercept.

The sub-area hierarchical Bayesian model is

$$\begin{aligned}\hat{\theta}_{ij}|\theta_{ij}, \hat{\sigma}_{ij}^2 &\overset{ind}{\sim} N(\theta_{ij}, \hat{\sigma}_{ij}^2),\ i = 1, \ldots, m,\\ \theta_{ij}|\boldsymbol{\beta}, \sigma_\mu^2 &\overset{ind}{\sim} N(\boldsymbol{x}_{ij}'\boldsymbol{\beta} + \nu_i, \sigma_\mu^2),\ j = 1, \ldots, n_i,\\ \nu_i|\sigma_\nu^2 &\overset{iid}{\sim} N(0, \sigma_\nu^2),\end{aligned}$$

where $(\boldsymbol{\beta}, \sigma_\mu^2, \sigma_\nu^2)$ is a set of nuisance parameters. The covariate $\boldsymbol{x}_{ij}$ is the maximum of FSA and RMA administrative data in county i within district j. The county-level FSA and RMA planted acreage data are highly correlated. To avoid the multicollinearity problem, we choose to use the maximum of these two data sources. Note that the above sub-area level model without sub-area level (ASD) effects, ν_i, reduces to the basic area level FH model without constraints.

A diffuse prior is adopted to the coefficients $\boldsymbol{\beta}$, that is, a multivariate normal prior distribution with fixed and known mean and variance and covariance matrix $\boldsymbol{\beta} \sim MN(\hat{\boldsymbol{\beta}}, 1000\hat{\Sigma}_{\hat{\boldsymbol{\beta}}})$. Here, $\hat{\boldsymbol{\beta}}$ are the least squares estimates of $\boldsymbol{\beta}$ obtained from fitting a simple linear regression model of the county-level survey estimates on the auxiliary data $\boldsymbol{x}_{ij}$ and $\hat{\Sigma}_{\hat{\boldsymbol{\beta}}}$ is the estimated covariance matrix of $\hat{\boldsymbol{\beta}}$. The prior distributions for σ_μ^2 and σ_ν^2 are Uniform $(0, 10^8)$ and Uniform $(0, 10^8)$. See the discussion in Browne and Draper (2006) and Gelman (2006) related to the prior of variance components in Bayesian models.

3.2 Models with Constraints

As discussed in Section 2.2, the constraints of the county-level estimates must be larger than the corresponding FSA and RMA planted acres data and the sum of all estimates within one state should be equal to the pre-published state-level estimate. In this section, the hierarchical Bayesian models with inequality constraints by Nandram, Cruze and Erciulescu (2020) are discussed.

First, the inequality constraints between the model estimates and administrative values need to be included in the model; that is,

$$\theta_{ij} \geq c_{ij}, i = 1, \ldots, m; j = 1, \ldots, n_i, \tag{1}$$

where the c_{ij} are fixed known quantities.

In our application on planted acres, $c_{ij} = max(\text{FSA}_{ij}, \text{RMA}_{ij})$ is defined as the maximum value between FSA and RMA corresponding values in the same county. Notice that in Figure 2, some of the survey estimates are one or two standard deviations below their corresponding c_{ij}, thereby creating some difficulties for the model estimates to do the same. The benchmarking constraint creates an additional challenge because the state target may be only slightly larger than the state total from administrative data, $c = \sum_{i=1}^{m}\sum_{j=1}^{n_i} c_{ij}$. This may be a tight condition as discussed in Cruze et al. (2019).

In addition, under NASS's top-down approach for benchmarking, the benchmarking constraint needs to be considered as well. In this paper, we are considering Bayesian models using Markov chain Monte Carlo (MCMC) simulation. After model fitting, a series of MCMC samples are obtained to construct the posterior summaries of interest. The ratio benchmarking adjustment method is adopted at the (MCMC) iteration level discussed in Erciulescu, Cruze and Nandram (2020) in the output analysis to address the county-state benchmarking constraint. It provides a suitable benchmarking adjustment to ensure consistency of county-level estimates with the state target efficiently.

Let θ_{ij}^B be the adjusted model estimates for county i and district j, $\theta_{ij,k}^B$ be the adjusted model estimates at the k^{th} iteration, and $\theta_{ij,k}$ be the model estimates at the k^{th} iteration, $k = 1, \ldots, K$. Let a be the state-level target.

The arithmetic mean of the MCMC samples is used to construct the point estimates of interest. After the ratio benchmarking adjustment,

$$\theta_{ij}^B = \frac{1}{K}\sum_{k=1}^{K}\theta_{ij,k}^B = \frac{1}{K}\sum_{k=1}^{K} r_k\theta_{ij,k}, \tag{2}$$

where r_k is the adjusted ratio at iterate level and the ratio r_k is

$$r_k = a \times \Big(\sum_{i=1}^{m}\sum_{j=1}^{n_i}\theta_{ij,k}\Big)^{-1}. \tag{3}$$

Therefore, the following relationship holds for county-state benchmarking

$$\begin{aligned}
\sum_{i=1}^{m}\sum_{j=1}^{n_i}\theta_{ij}^B &= \sum_{i=1}^{m}\sum_{j=1}^{n_i}\Big[\frac{1}{K}\sum_{k=1}^{K} r_k\theta_{ij,k}\Big] \\
&= \sum_{i=1}^{m}\sum_{j=1}^{n_i}\Big[\frac{1}{K}\sum_{k=1}^{K} a \times \Big(\sum_{i'=1}^{m}\sum_{j'=1}^{n_i}\theta_{i'j',k}\Big)^{-1}\theta_{ij,k}\Big] \\
&= a \times \frac{1}{K}\sum_{k=1}^{K}\Big[\Big(\sum_{i'=1}^{m}\sum_{j'=1}^{n_i}\theta_{i'j',k}\Big)^{-1}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\theta_{ij,k}\Big] \\
&= a.
\end{aligned}$$

However, we need to make sure the adjusted final estimate θ_{ij}^B can satisfy the inequality constraint as well. Given (1), the inequality constraint can be preserved for $\theta_{ij,k}$ in each k^{th} iteration. If $r_k \geq 1$ for each k, the following relationship follows from combining (1) and (2):

$$\theta_{ij}^B = \frac{1}{K}\sum_{k=1}^{K} r_k\theta_{ij,k} \geq \frac{1}{K}\sum_{k=1}^{K}\theta_{ij,k} \geq \frac{1}{K}\sum_{k=1}^{K} c_{ij} \geq c_{ij}. \tag{4}$$

Therefore, the inequality constraint $r_k \geq 1$; that is, $\sum_{i=1}^{m}\sum_{j=1}^{n_i}\theta_{ij,k} \leq a$ is needed for each k^{th} iteration because when the model estimates are raked up, they will satisfy the individual county's inequality constraints.

Based on the discussion above, θ_{ij} should be drawn subject to the constraints

$$\theta_{ij} \geq c_{ij}, i = 1, \ldots, m; j = 1, \ldots, n_i, \ \sum_{i=1}^{m}\sum_{j=1}^{n_i}\theta_{ij} \leq a \tag{5}$$

to address both inequality and benchmarking constraints in the models.

According to the constraints (5),

$$\sum_{i=1}^{m}\sum_{j=1}^{n_i} c_{ij} \leq \sum_{i=1}^{m}\sum_{j=1}^{n_i}\theta_{ij} \leq a.$$

Therefore, the support of θ_{ij} given $\underset{\sim}{\theta}_{(ij)}$ is

$$\max(c_{ij}, \sum_{i=1}^{m}\sum_{j=1}^{n_i} c_{ij} - \sum_{i'=1,i'\neq i}^{m}\sum_{j'=1,j'\neq j}^{n_i}\theta_{i'j'}) \leq \theta_{ij} \leq a - \sum_{i'=1,i'\neq i}^{m}\sum_{j'=1,j'\neq j}^{n_i}\theta_{i'j'}, \tag{6}$$

where $i = 1, \ldots, m; j = 1, \ldots, n_i$ and the lower bound $\mathbf{C} = (c_{ij})'$ are known and fixed.

To preserve the relationships after fitting the model, the constraint (6) is added to the FH model and the sub-area model in the priors to get the joint posterior density of $\theta_{ij}, i = 1, \ldots, m; j = 1, \ldots, n_i$. This problem falls under the general heading of constraint problems in statistics (e.g. Nandram, Sedransk and Smith 1997).

Therefore, the sub-area hierarchical Bayesian model with constraints is proposed as

$$\begin{aligned}
\hat{\theta}_{ij}|\theta_{ij}, \hat{\sigma}^2_{ij} &\overset{ind}{\sim} N(\theta_{ij},\ \hat{\sigma}^2_{ij}),\ j = 1, \ldots, n_i, \\
\theta_{ij}|\boldsymbol{\beta}, \delta^2 &\overset{ind}{\sim} N(\boldsymbol{x}'_{ij}\boldsymbol{\beta} + \nu_i, \sigma^2_\mu),\ \theta_{ij} \in \mathcal{T}, \\
\nu_i|\sigma^2_\nu &\overset{iid}{\sim} N(0, \sigma^2_\nu),\ i = 1, \ldots, m,
\end{aligned}$$

where $\mathcal{T}$ denotes the support (6) of θ_{ij} such that both the benchmarking constraint and the inequality constraints are satisfied simultaneously. $(\boldsymbol{\beta}, \sigma^2_\mu, \sigma^2_\nu)$ is a set of nuisance parameters and $\boldsymbol{x}_{ij} = (1, x_{ij1}, \ldots, x_{ijp})$ is the vector of covariates and the intercept. Note that the above sub-area level model without sub-area level (ASD) effects, ν_i, reduces to the area-level FH model with constraints. A diffuse prior is adopted to the coefficients $\boldsymbol{\beta}$, the same as the prior mentioned in Section 3.1. The non-informative prior distributions for σ^2_μ and σ^2_ν are Uniform $(0, 10^{10})$ and Uniform $(0, 10^{10})$, respectively.

It is worth noting that the state target should be equal to or greater than the administrative state total, $a \geq c$. That is, $a = \sum_{i=1}^{m} \sum_{j=1}^{n_i} \theta^B_{ij} \geq \sum_{i=1}^{m} \sum_{j=1}^{n_i} c_{ij} = c$. Therefore, there are feasible solutions to the inequality constraint problem in (5), and a feasible solution clearly depends on the target and the FSA and RMA values. As discussed in Section 2.2, most of the survey estimates are within two standard deviations of the bounds, but many of the smaller ones are much further below the bounds. Because of the advantage of shrinkage estimation in a small-area model, the smaller survey estimates are likely to be pulled upwards, and this will help to meet the bounds, but it does not solve the problem. If the model does not incorporate the inequality constraints, the final estimates do not necessarily cover the lower bounds in all cases. The inequality constraints need to be addressed in the models.

4. Case Study

The four models discussed in Section 3 are compared: the sub-area level model with inequality constraint, the area level model with inequality constraint, the sub-area level model without inequality constraint and the area-level model without inequality constraint. In addition, all models are fit using administrative data sources of information described in Section 2.2.

All models produce 2014 CAPS estimates of planted acres for corn in Illinois (IL) and Ohio (OH). FSA and RMA administrative data in IL usually have very high coverage rates of the planted acres for corn in each county. But in some specific counties in OH, both sets of administrative data have relatively low coverage rates for planted acres. The model performance is evaluated for both scenarios.

As mentioned in Section 2.2, the county-level survey estimates did not automatically cover all FSA and RMA administrative data. The relationship between survey estimates and the corresponding lower bounds based on administrative data (the maximum of FSA and RMA data) is displayed in Figure 4. The plotted pairs of survey estimates and administrative data are scattered around the 45 degree line. Around 31% of the county-level survey estimates cover FSA and RMA. About 56% of the survey estimates cover FSA and RMA for OH.

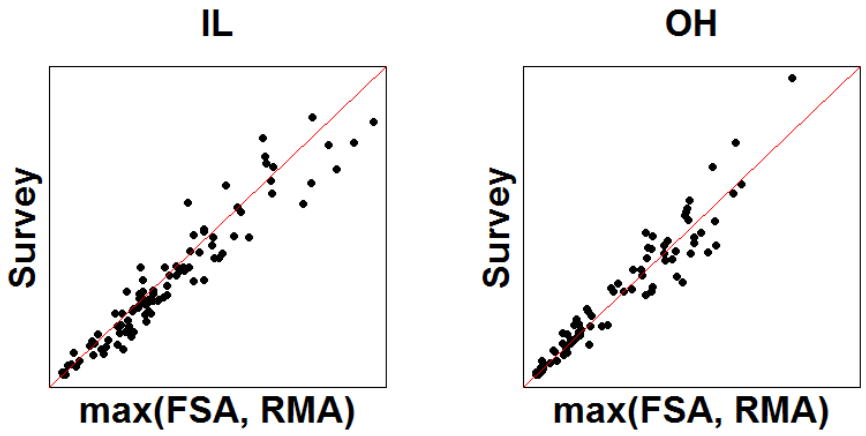

Figure 4: The County-level Planted Acreage Estimates of Corn for CAPS data and the lower bounds in IL and OH

In Section 4.1, a summary of the model fitting process is provided. Section 4.2 includes the internal checks for all four models. Several diagnostic tools are explored to check the adequacy of the models. External checks between model estimates, survey estimates and official statistics from NASS are presented for all models in Section 4.3.

4.1 Model Estimation

All four models are applied to all counties with positive data within one state for which $(\hat{\theta}_{ij}, \hat{\sigma}^2_{ij}, \boldsymbol{x}_{ij})$ are available. In IL, there are 102 counties and 9 ASDs in the CAPS samples for planted acreage. In OH, there are 88 counties and 9 ASDs.

MCMC simulation method is used to fit all four hierarchical Bayesian models using R and JAGS (Plummer, 2003). In each model, three chains are run for our MCMC simulation. Each chain contains 50,000 Monte Carlo samples, and the first 15,000 iterates are discarded as a burn-in to improve the mixing of each chain. After that point, 35,000 further iterations were produced for each of the three chains. In order to eliminate the correlations among neighboring iterations, those iterations are thinned by taking a systematic sample of 1 in every 35 samples. Finally 1,000 MCMC samples in each chain are obtained for constructing the posterior distributions of all the parameters, the nuisance parameters and the parameters for the planted acres.

The posterior means (PM) and posterior standard deviations (PSD) for parameters $\beta' = (\beta_0, \beta_1)$ and the variances, σ^2_μ and σ^2_ν for sub-area level models and σ^2_μ for area-level models for IL and OH, respectively, are displayed in Table 1-4. The signs of the coefficient β_1 in both states are as expected. The administrative data are very significant predictors with a positive sign for the county-level number of planted acres. For IL and OH, all the $\hat{\beta}_1$ are significant among four models but the intercept $\hat{\beta}_0$ are not significant (the 95% credible intervals contain zero) in both unconstrained models.

Model	Parameters	PM	PSD	ESS	$\hat{R}$
C Sub Area	β_0	-108676.1	56678.4	3000	1.001
	β_1	0.955	0.364	3000	1.002
	σ^2_μ	11765649.3	14368379.5	1900	1.012
	σ^2_ν	3.48E+09	2.76E+09	3000	1.003
NC Sub Area	β_0	-818.5	2111.3	3000	1.001
	β_1	0.884	0.024	3000	1.001
	σ^2_μ	17356714.6	15844279.3	2100	1.004
	σ^2_ν	10975427.3	17396487.7	3000	1.001

Table 1: Posterior means (PM), posterior standard deviations (PSD), effective sample sizes (ESS) and $\hat{R}$ for sub-area level models for 2014 IL corn

Convergence diagnostics are conducted. The convergence is monitored using trace plots, the multiple potential scale reduction factors ($\hat{R} \leq 1.05$) and the Geweke test of stationarity for each chain (Gelman and Rubin, 1992 and Geweke, 1992). Also, once the simulated chains have mixed, the effective number of independent simulation draws to monitor simulation accuracy is determined. Effective sample sizes and the $\hat{R}$ are shown in Tables 1-4, resulting in good convergence for all four models for both IL and OH. The values of $\hat{R}$ of most coefficient parameters are close to 1 and all of them are less than 1.05. The effective sample sizes of coefficient parameters in sub-area level models are 3000 and those in area-level models are around 2000 for IL. The effective sample sizes vary from 1100 to 3000 for OH.

Model	Parameters	PM	PSD	ESS	$\hat{R}$
C Area	β_0	-92856.5	53764.1	1500	1.010
	β_1	0.956	0.353	2000	1.002
	σ^2_μ	11430283	14932670	2100	1.008
NC Area	β_0	-558.4	1472.2	1700	1.002
	β_1	0.882	0.023	1900	1.002
	σ^2_μ	15096344.3	14654002.4	2500	1.004

Table 2: Posterior means (PM), posterior standard deviations (PSD), effective sample sizes (ESS) and $\hat{R}$ for area level models for 2014 IL corn

Model	Parameters	PM	PSD	ESS	$\hat{R}$
C Sub Area	β_0	-47706.4	25639.2	1800	1.006
	β_1	1.115	0.476	3000	1.002
	σ^2_μ	5329872.2	5464708.8	2000	1.007
	σ^2_ν	2.91E+09	2.83E+09	2300	1.003
NC Sub Area	β_0	-16.213	648.264	2800	1.004
	β_1	0.945	0.025	1400	1.006
	σ^2_μ	484029.4	538539.2	1200	1.007
	σ^2_ν	1774398.9	3515856.6	1800	1.003

Table 3: Posterior means (PM), posterior standard deviations (PSD), effective sample sizes (ESS) and $\hat{R}$ for sub-area level models for 2014 OH corn

Model	Parameters	PM	PSD	ESS	$\hat{R}$
C Area	β_0	-41862.9	23357.1	2300	1.007
	β_1	1.075	0.419	1700	1.009
	σ_μ^2	3895790.9	4085072.1	1900	1.008
NC Area	β_0	108.911	218.078	3000	1.001
	β_1	0.941	0.021	1100	1.017
	σ_μ^2	387476.8	431343.2	1200	1.007

Table 4: Posterior means (PM), posterior standard deviations (PSD), effective sample sizes (ESS) and $\hat{R}$ for area level models for 2014 OH corn

4.2 Internal Check

Several diagnostic tools are available to check the adequacy of all four models considered in this paper. First, the fit of models to data are assessed using posterior predictive checks (Rubin, 1984 , Meng, 1994). If the model fit is adequate to all observations $\hat{\theta}$, replicated values θ_{rep} that generated data from the model would be similar to observations. We calculate the Bayesian predictive p-value (BPP) to measure the adequacy of all models to the data from Gelman et al. (2004). The Bayesian posterior predictive p-value (BPP) is defined as

$$p = Pr\big(T(\theta^{rep}, \Omega) > T(\hat{\theta}, \Omega)|\hat{\theta}\big),$$

where $T(\theta, \Omega)$ is selected as $T(\theta, \Omega) = \sum_{i=1}^{m}\sum_{j=1}^{n_i} \frac{(\theta_{ij} - E(\theta_{ij}|\hat{\theta}))^2}{Var(\theta_{ij}|\Omega)}$ and Ω are the nuisance parameters in each model. The p-value is the probability of the sum of square residuals based on replicated estimates larger than the one from observed data. If the value is extreme (close to 0 or 1), it indicates a discrepancy between the model and the data, meaning the model is not adequate. The BPP for each model is presented in Table 5. For IL, the BPPs in area level and sub-area level models with constraints are 0.663 and 0.504, respectively, which are close to 0.5. The models without constraints have high BPP, 0.903 and 0.947, respectively, but they are not extremely close to 1 (e.g. close to 0.99). Similar results show for OH in Table 5. These BPPs did not raise the concerns on all four models based on data for IL and OH.

Another goodness-of-fit measure for all four models is the deviance information criterion (DIC) (Spiegelhalter et al., 2002) shown in Table 5. The DICs from sub-area models are slightly smaller than those in area-level constrained models. The DICs of unconstrained models are smaller than those in constrained models. Usually the model with the smallest DIC is selected to be the model that would best predict a replicate dataset that has the same structure as that currently observed. The sub-area level models are better than the area level models.

Based on the DICs, the unconstrained models tend to produce predictions close to the observed data. The constrained models produce replicate datasets based on the constraints, and they are not necessarily similar to the observed values. However, it is not quite suitable to make the model selection only based on DICs in this case between constrained and unconstrained models. Note that the scope of this paper is to generate county-level model estimates that satisfy all required constraints rather than making predictions based on observed data and administrative data. Figures 2 and 4 show the scatter plots between observed data and administrative data. The observed data are not necessarily above the corresponding administrative data. The inequality constraints are introduced into the models, and they produce more "biased" model-based estimates of the observed data when compared to the models without inequality constraints. The bias correction terms in DIC tend

to be larger. Therefore, it is hard to make a decision based on BPP and DIC diagnostics between sub-area level constrained and unconstrained models. To check model performance between sub-area level constrained and unconstrained models, external comparisons are shown in the next section.

Type	Model	DIC		BPP	
		C	NC	C	NC
IL	Sub-area	2334.6	2285.3	0.504	0.947
	Area	2335.7	2285.2	0.633	0.903
OH	Sub-area	1881.1	1766.7	0.331	0.967
	Area	1884.7	1776.4	0.248	0.908

Table 5: DICs and BPPs for constrained and unconstrained models

4.3 External Check

Internal checks show that all sub-area level models provide adequate fit to the data and sub-area level models have slightly smaller DICs. However, none of the internal checks considered reveal much in terms of the model performance of both sub-area level constrained and unconstrained models. In this section, the inequality constraints check is conducted and the estimates for each model is compared with the published estimates.

First, the inequality check between the final model estimates number of planted acres and the corresponding FSA and RMA administrative data is conducted for each model. See Figures 5 and 6. Counties in light color indicate that the corresponding model estimates are smaller than FSA and RMA data. Counties in dark color mean that their estimates are larger than the maximum of both FSA and RMA administrative data. The right maps in both Figures 5 and 6 show the results of constrained model, indicating that all counties in both states cover the administrative data after ratio benchmarking.

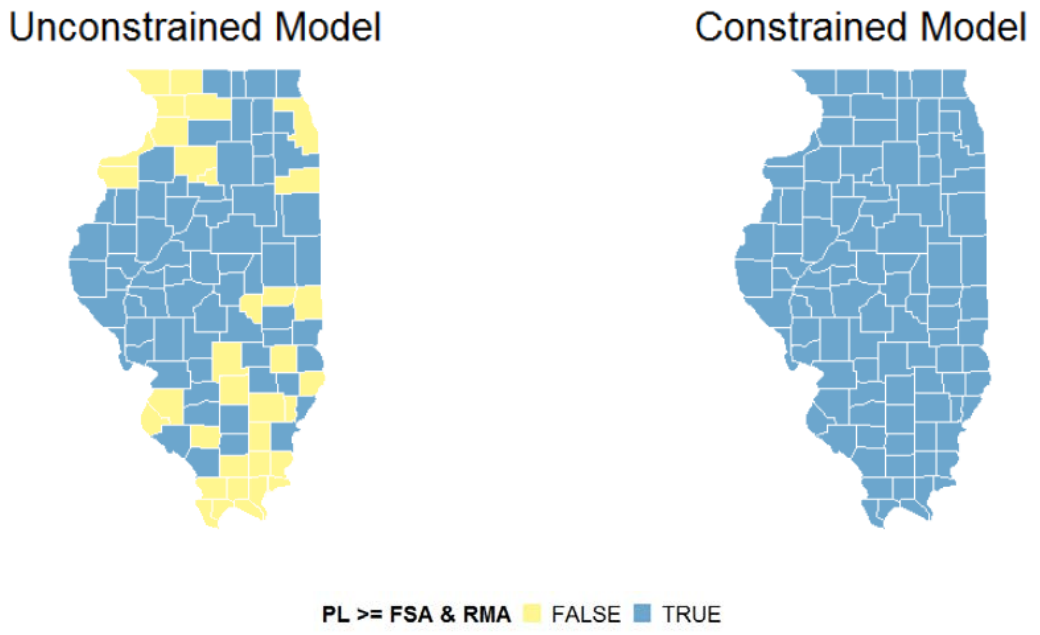

Figure 5: Inequality check for unconstrained and constrained models for IL

Approximately 75% of the unconstrained model estimates covered the floor imposed by administrative data. *Some counties lost more than 1,000 planted acres.* (Note that there

are 640 acres to the square mile.) For OH, approximately 91% of the unconstrained model estimates cover the administrative acreage data. Therefore, for the model without inequality constraints, the coverage rate on administrative data depends on the relationship between survey estimates and the administrative data. Those models cannot promise to produce model-based (ME) estimates that cover all administrative data.

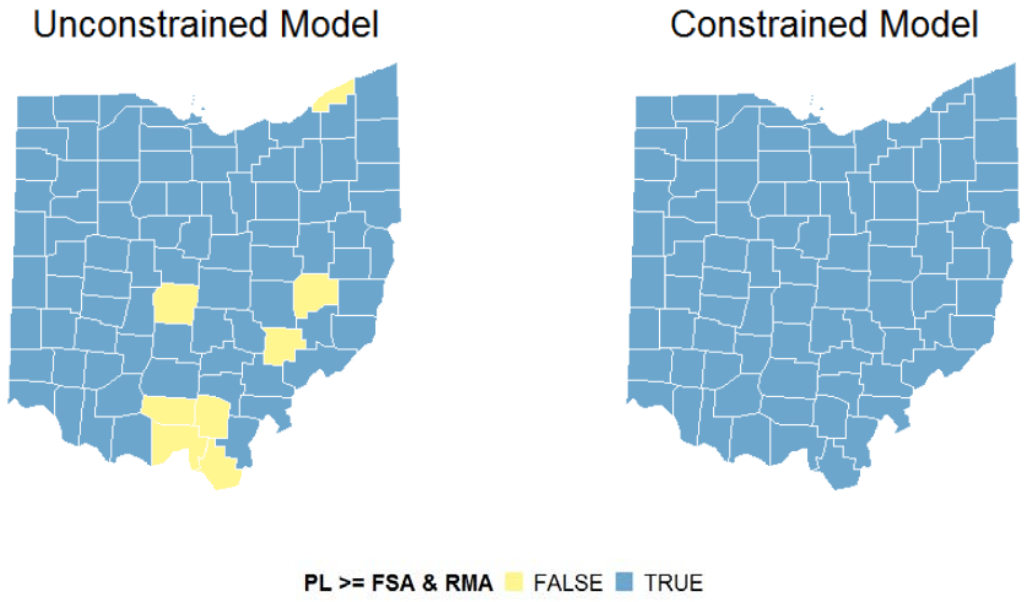

Figure 6: Inequality check for unconstrained and constrained models for OH

In addition, unconstrained model (NC) estimates, constrained model (C) estimates and the survey (DE) estimates are compared with the published estimates. The absolute relative differences between different estimates and published estimates,

$$\text{ARD} = \frac{|\theta_{MERB} - \text{Published}|}{\text{Published}},$$

are calculated and presented, where θ_{MERB} is denoted as the final model estimates after ratio benchmarking. A small ARD is one key check on the performance of model-based point estimates.

The posterior coefficients of variation (CV),

$$\text{CV} = \frac{\text{PSD}}{\theta_{MERB}},$$

are calculated, where PSD is the corresponding posterior standard deviation of θ_{MERB} from different models (see Table 6 and Table 7).

The number of responses for number of planted acres in CAPS varies with county in each state. Many counties in IL and OH have relatively large number of responses. However, many counties have only few responses. Small area models tend to improve the accuracy of estimates comparing to the accuracy of survey estimates, especially in areas with small sample sizes. In order to examine the effect of sample size among our models, we split counties of IL and OH, respectively, into three groups according to their number of reports in CAPS: small sizes (less than 30); median sizes (between 30 and 60); large sizes (larger than 60). We showed all statistics in Table 6 and Table 7 as well.

Among all counties in IL, the median ARD value between survey estimates and published estimates in IL is 12.942%. Substantial improvement can be noticed from both the

constrained model and the unconstrained model. Again compared to published estimates, the median ARD value based on the constrained model is 0.194%, less than the median ARD value based on the unconstrained model, 0.948%. Moreover, the range of ARD values from the constrained model (0.003%, 34.908%) are much narrower than the range based on survey estimates (0.259%, 82.973%) and also less than those from the unconstrained model (0.007%, 51.349%). Therefore, for IL, the sub-area level model with constraints performs best among the unconstrained model and survey estimates as measured by the ARD. In addition, Table 6 shows the ARD values based on the sample sizes of counties in IL. The ranges of ARD values based on both models are large for counties with small number of reports. ARD values from the constrained model are within 2% for median size counties but those from the unconstrained model are from 0.007% to 17.036 %. For large counties, the relative differences from all models are the narrowest among all three types of counties. They are within 2% difference for constrained models and 3% from unconstrained model. As is to be expected, all estimates are closer to the published estimates with increasing sample size. Overall, the comparisons of ARD values show that the constrained model increases the accuracy of the estimates significantly.

The CVs of the IL modeled and survey estimates are shown in Table 6. The sub-area level model can borrow information both from covariates and from other counties within the district (sub-area) level. Therefore, the posterior CVs would have a greater reduction compared with the CVs of the survey estimates. The median CVs among all counties in IL are in decreasing order: survey, the unconstrained model and the constrained model. In the unconstrained model, the CVs of small size counties are the largest (20.544%, 125.905%). The maximum estimated CVs exceeds that in survey estimates. The CVs of the constrained model are much smaller than those from survey and the unconstrained model. As expected, the CVs are smaller when sample sizes increase. In the model with inequality constraints, the maximum CVs is in the small size counties as well.

Sample size	Statistics	ARD (%)			CV (%)		
		DE	NC	C	DE	NC	C
Overall	Min	0.259	0.007	0.003	10.501	1.899	0.144
	Median	14.914	0.948	0.194	19.210	5.199	0.272
	Max	82.973	51.346	34.908	92.283	125.905	12.705
[0,30)	Min	0.259	0.622	0.273	25.315	20.544	1.466
	Median	16.585	13.530	0.978	42.421	34.905	2.187
	Max	66.174	51.346	34.908	92.283	125.905	12.705
[30,60)	Min	0.575	0.007	0.007	10.501	2.459	0.185
	Median	9.721	1.204	0.176	19.885	5.812	0.278
	Max	39.620	17.036	1.940	33.961	21.985	2.336
≥ 60	Min	7.474	0.096	0.003	9.108	1.899	0.144
	Median	33.990	0.646	0.196	15.731	3.151	0.214
	Max	82.973	2.032	1.199	53.570	5.522	1.740

Table 6: 2014 IL corn planted acres: comparisons of ARDs and CVs among survey, sub-area unconstrained model and constrained model

Table 7 shows all the comparisons for OH. The median of ARDs between survey estimates and published estimates is 12.942%. Substantial improvement can also be noticed from both constrained and unconstrained models. The median ARD value between model-based estimates and the published estimates is around 2%. The smallest median of the relative differences is 2.394% in the unconstrained model. However, the range of ARD values from the constrained model is (0.093%, 49.858%), which is narrower than the one from the unconstrained model, (0.103%, 95.376%). Notice that the ranges of ARDs in OH are larger than those in IL. The administrative data for OH are not stronger comparing with

those in IL. In several counties, FSA and RMA administrative data have the undercoverage issue.

To examine the effect of sample size, OH is split into three groups and all statistics are presented in Table 7. The ranges of the ARD values based on models and the survey are relatively large in small size counties. Both model estimates are much closer to the published estimates. The model estimates based on the constrained model in small size counties are closest to the published estimates based on the range of the ASD values. However, the median ARD value from the constrained model is 1.241 $\%$ for large size counties, which is larger than the one from the unconstrained model, 0.876%. The maximum ASD value is similar as well. In the median size counties, constrained model tends to provide larger estimates compared with those from unconstrained model. If there was no inequality constraint, the model estimates would be affected by the undercoverage issue from the administrative data when borrowing information from them.

The CVs are compared among models and the survey estimates for OH as well. Similar to IL, the posterior CVs based on the models have large reductions comparing with the CVs from survey. The median CV in the unconstrained model is 3.67$\%$, larger than the one in the constrained model. The maximum CV in the unconstrained model is the highest among models and survey. As expected, the CVs are smaller when sample sizes increase. The maximum of CVs is in small size counties as well. The CVs based on constrained model are much smaller than those of constrained model and survey. For OH, the range of CVs in model with inequality constraints are wider than those for IL.

Sample size	Statistics	ARD (%)			CV (%)		
		DE	NC	C	DE	NC	C
Overall	Min	0.002	0.103	0.093	8.754	1.043	0.473
	Median	12.942	2.394	2.575	22.292	3.670	0.797
	Max	114.123	95.376	49.858	100.000	104.411	89.816
[0,30)	Min	0.002	0.103	0.671	17.169	3.266	0.533
	Median	24.898	9.791	4.650	35.280	22.292	5.044
	Max	95.687	95.376	49.858	100.000	104.411	89.816
[30,60)	Min	1.574	0.136	0.093	10.224	1.206	0.473
	Median	12.699	2.266	2.191	19.468	2.546	0.660
	Max	114.123	10.968	14.864	33.072	29.994	10.548
≥ 60	Min	6.172	0.322	0.216	8.755	1.043	0.499
	Median	11.982	0.876	1.241	14.699	1.507	0.765
	Max	18.915	5.001	6.785	19.384	5.231	4.136

Table 7: 2014 OH corn planted acres: comparisons of ARDs and CVs among survey, sub-area unconstrained model and constrained model

5. Conclusion

NASS puts extensive research efforts on crops county estimate program aimed primarily to improve the precision of the estimates at county level while preserving the underlying relationships among the estimates and administrative data. Different small area estimation models are implemented to integrate multiple sources of auxiliary information with CAPS data. In this paper, models with inequality constraints are discussed and implemented to address the needs and challenges of the inequality and benchmarking constraints that NASS official statistics need to satisfy. That is, the county-level estimates of planted acreage should "cover" the corresponding administrative data while the total acreage of all available county-level estimates are equal to the state target.

We apply both sub-area and area-level models with inequality constraint to construct reliable and coherent county-level planted acreage estimates. In the case study of 2014 corn based on IL and OH, we show model diagnostics and provide internal checks among all four models. The internal checks show the sub-area level models are slightly better than area-level model. However, the residual-type internal checks are not very suitable for comparing the constrained and unconstrained model since our focus is to provide coherent estimates close to the official estimates but not to the observed data.

Now more comparisons among both sub-area level model estimates and survey estimates are made. First, the inequality checks show that constrained model can preserve the relationships among estimates and administrative data. But this is not necessarily the case for the unconstrained model. In addition, the statistics of ARD values show that the constrained model provides estimates closer to the published values than those from the unconstrained model as well as those from the survey, especially for IL. FSA and RMA are very significant covariates for the estimates of planted acres. Moreover, the associated measures of uncertainty (CVs) from models are significantly smaller than the CVs of the survey estimates. The basic sub-area models can reduce the CVs while borrowing strength from auxiliary information and all counties within one district and all districts within one state. In addition, for the constrained model, the prior information based on the lower bound information from FSA and RMA data and the upper bound related to the state target reduce the CVs of the model-based estimates since estimates can be drawn only in the restricted support. Therefore, the performance of the sub-area level model with inequality constraints illustrates significant improvement of county-level estimates of planted acres in accuracy and precision.

Major ongoing and future research related to sub-area level constrained model involves the investigation of different auxiliary information. The auxiliary information considered here is the key data sources of planted acres (the combination of FSA and RMA administrative data). Future efforts will be on searching and applying other useful data sources to strengthen the model. Remote sensing data, NASS cropland data layer (CDL), and weekly weather data are available at the county level. Variable selections should be investigated for different states and commodities because weather conditions influence the planting progress and the planted acres within different time periods based on different states and commodities.

In addition, missing data problems are another challenge for the application of the constrained model. In this paper, two case studies related to corn-belt states, IL and OH, which do not have missing data in 2014 corn, are provided. However, it is not always the case for other states or other commodities. As mentioned in Section 2.1, CAPS is conducted for different commodities among all eligible states. In some cases, the survey may not indicate any planted area with respect to a particular commodity, but administrative data might represent some positive acres or vise versa. Erciulescu, Nathan and Nandram (2020) uses the nearest neighbor methods to impute missing data for either survey or covariates. This approach and imputing and borrowing information from previous year or the average of several years estimates are being explored. How to deal with missing data and provide reliable and coherent predictions are ongoing research.

Disclaimer and Acknowledgment

The findings and conclusions of this paper are those of the authors and should not be construed to represent any official USDA or U.S. Government determination or policy. This research was supported in part by the intramural research program of the U.S. Department of Agriculture, National Agriculture Statistics Service.

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.

Browne, W. J. and Draper, D. (2006). A comparison of bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Analysis*, 1(3):473–514.

Cruze, N. B., Erciulescu, A. L., Nandram, B., Barboza, W. J., and Young, L. J. (2019). Producing official county-level agricultural estimates in the united states: Needs and challenges. *Statistical Science*, 34(2):301–316.

Erciulescu, A. L., Cruze, N. B., and Nandram, B. (2019). Model-based county level crop estimates incorporating auxiliary sources of information. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 182(1):283–303.

Erciulescu, A. L., Cruze, N. B., and Nandram, B. (2020). Statistical challenges in combining survey and auxiliary data to produce official statistics. *Journal of Official Statistics*, 36(1):63–88.

Fay III, 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(366a):269–277.

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

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper). *Bayesian Analysis*, 1(3):515–534.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical science*, 7(4):457–472.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In *Bayesian Statistics 4*. Citeseer.

Meng, X.-L. (1994). Posterior predictive p-values. *The Annals of Statistics*, 22(3):1142–1160.

Nandram, B., Sedransk, J., and Smith, S. J. (1997). Order-restricted bayesian estimation of the age composition of a population of atlantic cod. *Journal of the American Statistical Association*, 92(437):33–40.

Nandram, B., Cruze, N. B., and Erciulescu, A. L. (2020). Bayesian benchmarking under inequality constraints and double shrinkage. *Working Paper*, NASS, USDA.

National Academies of Sciences, Engineering, and Medicine (2018). *Improving crop estimates by integrating multiple data sources*. National Academies Press.

Pfeffermann, D. (2013). New important developments in small area estimation. *Statistical Science*, 28(1):40–68.

Rao, J. N. K. and Molina, I. (2015). *Small area estimation*. 2015 John Wiley and Sons, Inc.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applies statistician. *The Annals of Statistics*, pages 1151–1172.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the royal statistical society: Series B (statistical methodology)*, 64(4):583–639.

Torabi, M. and Rao, J. (2014). On small area estimation under a sub-area level model. *Journal of Multivariate Analysis*, 127:36–55.

U.S. Department of Agriculture Office of the Chief Economist (2019). Update of 2019 FSA Acreage Data and FAQs on USDA Acreage. page 3.