Combining survey and administrative data to produce official statistics

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

Abstract

Combining survey and auxiliary data to produce official statistics is gaining interest at federal agencies and among policy makers due to its efficiency. Moreover, there is an increase in reliable estimates at detailed, disaggregated levels, a decrease in allocated budgets and an increase in usability of available data. The United States Department of Agriculture's National Agricultural Statistics Service issues annually nearly 500 reports, including county-level estimates that play an important role in the allocation of funds in some agricultural programs. In this paper, small area estimation modeling approaches are considered that borrow information across areas and from auxiliary data, to produce reliable county-level agricultural predictions. Challenges in assessing the quality and the usability of different data sources are discussed in the context of planted acreage estimation.

Key Words: Administrative Data, End-of-Season Agricultural Quantities, Official Statistics, Small Area Estimation

1. Introduction

Survey summary statistics at disaggregated levels may not be fit for use as official statistics because the limited amount of information available may result in estimates with high levels of uncertainty. With an increase in available data from auxiliary sources, an increase in needs for official statistics at detailed levels of aggregation and a decrease in allocated budgets, federal agencies have an increased interest in using models in the estimation process. For example, the United States Department of Agriculture's (USDA's) National Agricultural Statistics Service (NASS) conducts an annual series of surveys to provide acreage, production and yield estimates at state and substate levels of aggregation. The survey summary statistics are used to produce the final official statistics, which play an important role in the policy and decision making by other USDA agencies, such as the Farm Service Agency (FSA) and the Risk Management Agency (RMA). In this paper, we explore auxiliary data sources and model-assisted methods to produce predictions for counties with survey sample sizes as small as zero.

Area-level and subarea-level models are excellent reproducible tools that combine survey data and auxiliary data to produce reliable estimates for areas where survey estimates are available. In the area-level model introduced by Fay and Herriot in 1979 (FH), the survey estimates, $\hat{\theta}_k$, are modeled using the sampling model,

$$\hat{\theta}_k|(\theta_k, \hat{\sigma}_k^2) \overset{ind}{\sim} N(\theta_k, \hat{\sigma}_k^2),$$

where $\hat{\sigma}_k^2$ are the estimated sampling variances and $k = 1, ..., m$ is an index for the small area. The small area parameter of interest θ_k is estimated using a linkage model,

$$\theta_k|(\boldsymbol{\beta}, \sigma_u^2) \overset{ind}{\sim} N(\mathbf{x}_k'\boldsymbol{\beta}, \sigma_u^2), \tag{1}$$

[*] National Institute of Statistical Sciences and USDA National Agricultural Statistics Service, 1400 Independence Avenue, SW, Room 6412B, Washington, DC 20250-2054. E-mail: aerciulescu@niss.org.

[†] USDA National Agricultural Statistics Service, 1400 Independence Avenue, SW, 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.

where $\mathbf{x}_k$ are area-level covariates with p components, including an intercept, and $(\boldsymbol{\beta}, \sigma_u^2)$ is a vector of nuisance parameters. A rich literature is available for the FH model and its extensions, using both frequentist and Bayesian methods. In a hierarchical Bayes analysis, prior distributions are assigned to $(\boldsymbol{\beta}, \sigma_u^2)$.

As an extension to the FH model, Fuller and Goyeneche (1998) introduced a subarea-level model (FG), to account for a grouping structure of the small subareas into areas. The survey estimates at the subarea level, $\hat{\theta}_{ij}$, are modeled using the sampling model,

$$\hat{\theta}_{ij}|(\theta_{ij}, \hat{\sigma}_{ij}^2) \overset{ind}{\sim} N(\theta_{ij}, \hat{\sigma}_{ij}^2),$$

where $\hat{\sigma}_{ij}^2$ are the estimated sampling variances, $j = 1, ..., n_i^c$ is an index for the small subareas, $i = 1, ..., m$ is an index for the areas, and $n^c = \sum_{i=1}^{m} n_i^c$ is the total number of subareas. The parameter of interest is the small subarea mean θ_{ij}, which is estimated using a hierarchical linkage model,

$$\begin{aligned} \theta_{ij}|(\boldsymbol{\beta}, \sigma_u^2, v_i) &\overset{ind}{\sim} N(\mathbf{x}_{ij}'\boldsymbol{\beta} + v_i, \sigma_u^2), \\ v_i|\sigma_v^2 &\overset{ind}{\sim} N(0, \sigma_v^2), \end{aligned} \tag{2}$$

where $\mathbf{x}_{ij}$ are subarea-level covariates with p components, including an intercept, and $(\boldsymbol{\beta}, \sigma_u^2, \sigma_v^2)$ is a vector of nuisance parameters. Torabi and Rao (2014) studied the FG model in a frequentist framework and Kim et al. (2018) extended the linkage model in Torabi and Rao (2014) to allow for a hierarchical level for parameters $\boldsymbol{\beta}$ and to remove distributional assumptions in the first hierarchical level. Erciulescu et al. (2016, 2017, 2018) studied the FG model using a hierarchical Bayes framework, adopting prior distributions for $(\boldsymbol{\beta}, \sigma_u^2, \sigma_v^2)$.

In the area-level (subarea-level) sampling models, it is assumed that $\hat{\theta}_k(\hat{\theta}_{ij})$ and $\hat{\sigma}_k^2(\hat{\sigma}_{ij}^2)$ are valid estimates available from the survey summary, i.e., the estimates exist and are in the parameter space. However, for the not-in-sample subareas (domains with missing survey data), inference conducted relies on the linkage model's specification. Given (1), a typical choice of estimator for the not-in-sample areas is the synthetic estimator $\mathbf{x}_k'\boldsymbol{\beta}$, see Rao and Molina (2015) for more information on regression synthetic estimation. While one choice for a not-in-sample subarea estimator, given (2), is the synthetic estimator $\mathbf{x}_{ij}'\boldsymbol{\beta}$, a better estimator is the composite estimator $\mathbf{x}_{ij}'\boldsymbol{\beta} + v_i$ (note the contribution of both the subarea-level auxiliary data and the area-level random effect). In a Bayesian approach, the predictions are drawn from the assumed linkage model (1) or (2), for area-level or subarea-level, respectively.

In this paper, we consider data collected by the USDA's NASS using a probability sample and auxiliary data from other sources, to produce end-of-season county-level and agricultural statistics district-level predictions for planted acreage, where an agricultural statistics district (hereafter, denoted by ASD or district) is defined as a group of contiguous counties within a state. In particular, the probability sample of interest to this study is the pooled sample from the quarterly crops Agricultural Production Surveys (USDA NASS APS 2018) and their supplement, the County Agricultural Production Surveys (USDA NASS CAPS 2018), and will be denoted hereafter by CAPS.

Statistical challenges in combining data from multiple sources to produce official statistics are discussed throughout the paper. In Section 2, we introduce different data sources and present a method that combines survey data and administrative data to identify and predict planted acreage for in-sample and not-in-sample subareas of interest for certain crop, i.e., county-level corn, as in the case study illustrated here. Modeling strategies addressing different scenarios of available data and the corresponding derived predictors are presented in Section 3. In Section 4, we present nationwide prediction results for 2015 corn planted acreage, including model efficiency and different contributions of administrative data to produce official statistics. Concluding remarks are given in Section 5. Technical details on derivations of closed-form expressions for the model predictions are given in Appendix A. Additional results on corn, soybean, sorghum and winter wheat are presented in Appendices B and C.

2. Data for Modeling End-of-Season Crop Acreage

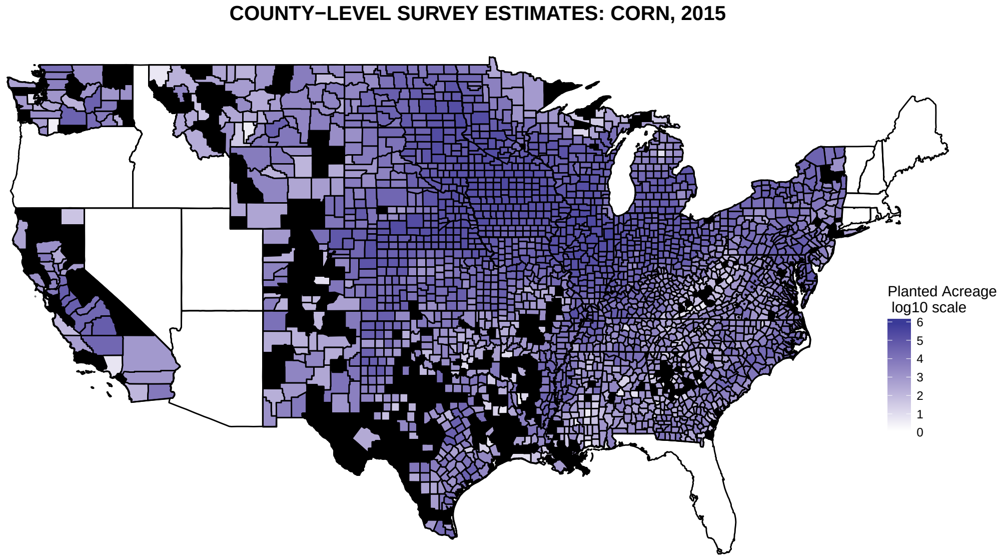
County-level survey estimates may be improved using auxiliary information and small area model-based procedures, especially for counties with small sample sizes. Estimation challenges are driven by the needs for multi-stage (county, district, state), nationwide, estimates, constructed using a small amount of survey data. In this section, we describe the sources of data considered to produce small area model predictions for end-of-season crop planted acreage for corn in 2015. Prediction is conducted state by state and commodity by commodity. The NASS survey data and the auxiliary data available from other USDA agencies, on corn planted acreage, are combined at the county level, for each state.

Due to the updates to the list sampling frame and the survey questionnaires, and to the year-to-year changes in planting activity, the set of subareas to be estimated for a given year-commodity combination is not predefined. For example, each survey response includes information on the entire operation (farm or ranch), and for all the sampled commodities with activity in the given season. As a result, the number of known operations in a county may change over time, the number of sampled operations may vary from year to year, and each of those operations may vary the type of crops grown annually. See Appendix A in National Academies of Sciences, Engineering, and Medicine (2017) for more details on NASS's survey design and data collection. In this section, we introduce a method that combines survey data and administrative data to identify the 2015 not-in-sample counties of interest for corn planted acreage prediction. Also, we investigate the potential for using auxiliary data as covariates in hierarchical models.

2.1 NASS Survey Data

County-level and ASD-level survey estimates and associated variance estimates are available from the NASS's CAPS summary. The ASD-level survey data are derived directly from the county-level survey data and, hence, only the county-level data will be used for modeling. The ASD-level survey data will be used for comparing model predictions to the survey estimates. In the 2015 crop season, NASS sampled 36 states for corn. The 36 states were comprised of 2837 counties, and NASS produced survey estimates for 2426 in-sample counties. Survey estimates are not available for the rest of 411 counties; we refer to these counties as not-in-sample with respect to corn. A nationwide map of the end-of-season positive county-level planted acreage survey estimates available for corn in 2015 is shown in Figure 1. The 12 states that were not sampled for corn in 2015 are represented as blank states. The counties with zero planted acreage predictions and not-in-sample counties for corn in 2015 are represented in black. Since the range of planted acreages in counties with available sample data is state-dependent and can vary from tens to hundreds of thousands of acres, the county-level map in Figure 1 depicts estimates on the log(10) scale. Dark-purple areas correspond to high acreage intensity regions, in particular the Midwestern corn belt states.

Figure 1

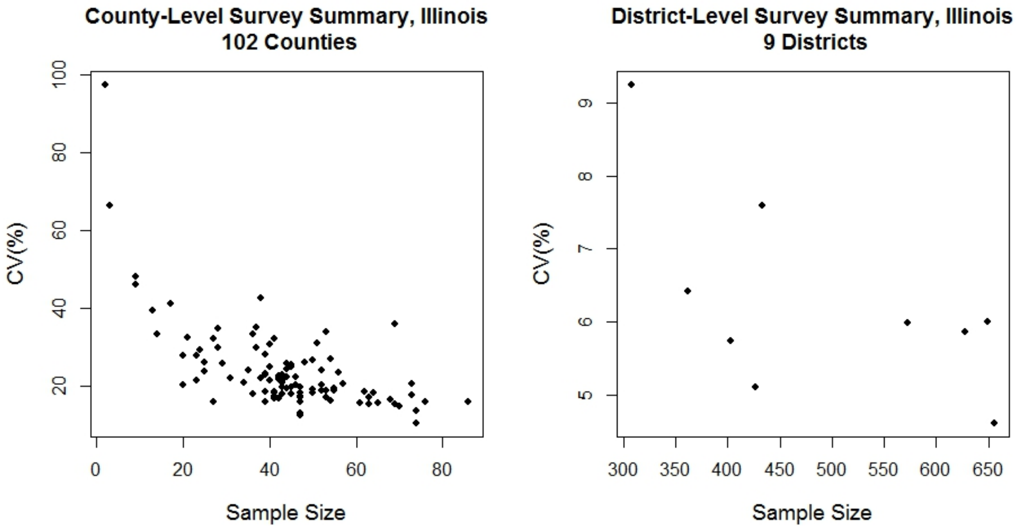


As a result of the NASS survey and publication cycle, state-level planted acreage values are prepublished and considered as fixed targets in the substate-level estimation process. The sum of the county-level survey estimates in a state does not necessarily equal the prepublished state-level value, the latter being the result of an expert assessment of multiple sources of data (including, but not limited to the survey data). Hence, one of the challenges encountered was to attain consistency among estimates constructed for nested levels. To overcome this challenge, we study a benchmarking adjustment applied to the substate-level predictions, for the county-to-ASD-to-state agreement to hold. More details on the benchmarking adjustment we utilize are presented in Section 3.3.

The number of counties and ASDs vary across the states and across commodities. For 2015 corn, the number of counties within ASDs ranges from 1 to 32, with a median of 8 and the number of ASDs within state ranges from 3 to 15, with a median of 9. Because the source of survey data for this study is the survey summary at the county level and ASD level, we denote the sample size by the number of positive records used to construct the survey summary; a positive record refers to a survey record for which positive acreage was reported. The county sample size differs from state to state and commodity to commodity. For 2015 corn, the county sample sizes range from 1 to 191, with a median of 18 and the district sample sizes range from 1 to 993, with a median of 206.

The estimated coefficients of variation (CVs) for the survey estimates increase as the county sample sizes decrease, and their ranges also differ from state to state and commodity to commodity. For 2015 corn, the CVs of the county-level survey estimates range from 0.07% to 107.66%, with a median of 31.94%, and the CVs of the ASD-level survey estimates range from 3.27% to 100.70%, with a median of 10.67%. Figure 2 shows the inverse relationship between the CVs of the 2015 corn county-level planted acreages survey estimates in Illinois and the corresponding sample sizes. Similar patterns are observed in other states, and for other commodities.

Figure 2



2.2 Auxiliary Data

We explore auxiliary data, available from three USDA agencies: FSA, RMA, and NASS. FSA administers U.S. farm programs, including those authorized by the Agricultural Act of 2014, known as the "Farm Bill" (USDA FSA 2014), e.g., county-level revenue loss protections. RMA oversees the Federal Crop Insurance Corporation, which provides crop insurance to participating farmers and agricultural entities (USDA RMA 2014). For this, FSA and RMA collect data from farmers participating in such programs. NASS produces the Cropland Data Layer (CDL 2018), a crop-specific land cover product that uses satellite and FSA ground-reference data to classify crop types on the continental United States (Boryan 2011, USDA NASS 2016a).

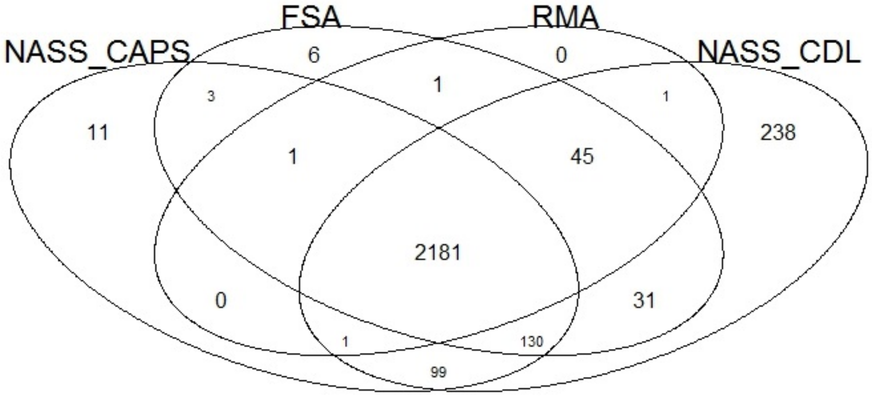
The levels and time of availability, and potential sources of error vary by data source (FSA, RMA, NASS), geography and commodity. Combining data from multiple sources and assessing its quality and usability is a challenging effort, often not mentioned in small area studies. For example, the CAPS sample data are collected on farms or ranches that the respondents operate and participation in the FSA and RMA programs is popular, but not compulsory; farmers who choose to participate in either agency's support programs supply data to the FSA and RMA administrative offices voluntarily. However, the definition of farm or ranch and the spatial unit used differ among the three data sources: NASS, FSA and RMA (National Academies of Sciences, Engineering, and Medicine, 2017, pages 96-97). Linking data at a fine scale has been of interest to NASS, but final solutions have yet to be developed. The administrative data of interest for this study are the self-reported corn planted acreage values supplied to FSA and RMA and the acreage values derived from pixels classified as corn, aggregated at the county level.

Quantifying the quality of nonprobability sample data has been of interest to many government agencies but conclusive studies have yet to be published. Parsons (1996) evaluated the quality of FSA acreage totals with respect to coverage. Kennedy et. al (2016) evaluated nonprobability surveys and assumed that the nonprobability samples were drawn as simple random samples from the population and constructed pseudo-weights when constructing domain estimates and associated measures of uncertainty. While we acknowledge potential error sources in the aggregated data, in this study we will treat the nonprobability county-level values as fixed and free of error. In Table 1, we report a summary of the number of counties with data available on corn planted acreage in 2015 from at least one source. Note that the sets of counties with data available from either of the four sources are not mutually exclusive, as depicted in the Venn diagram in Figure 3. After accounting for the 2726 counties with corn planted acreage identified from the CDL, additional planted acreage activity is identified in only 22 (= 11+3+6+0+1+1+0) counties from the CAPS, FSA and RMA, see Figure 3. Hence, our goal is to construct 2015 corn predictions for the total of 2748 counties. The number of counties with corn planting activity differs across years, states, commodities and data sources.

Table 1: Counties, *in Sampled States*, with Corn Planting Activity, 2015

Data Source (USDA)	Number of Counties
NASS CAPS	2426
FSA	2398
RMA	2230
NASS CDL	2726

Figure 3: Counties, *in Sampled States*, with Corn Planting Activity, 2015



Since the values available from the three sources of auxiliary data are measurements of the same county-level quantity, i.e., corn planted acreage, the three sources may be combined at the county level to construct one set of values indicating the maximum number of available, reported by volunteers or remotely classified, corn planted acreages. Let Admin PL denote the constructed variable as such. If all FSA, RMA and CDL values are available, then the maximum value is considered. If only two of the values are available, the maximum value is considered. If only one of the values is available, then that value is considered. To investigate the additional contributions of the CDL data, we will also consider an Admin PL variable, as derived from FSA and RMA data only, and present results in Section 4.

2.3 Borrow Information from Multiple Data Sources

Nationwide analysis indicates strong linear relationships between the survey estimates and the administrative data for all the states. A simple regression model of survey estimates on FSA, RMA, CDL or Admin PL administrative values produces R^2 values and estimated slope coefficients $\hat{b}$ summarized in Table 2 $(25\%, 50\%, 75\%$ quantiles). For illustration, we use the data available for all the 102 counties in Illinois, since planted acreage values are available from all the data sources. In Figure 4 we display the linear fit between the survey estimates and the derived administrative values, Admin PL, and in Figure 9 in Appendix C we display the linear fits between the survey estimates and the values available from each of the three auxiliary sources, FSA, RMA and CDL, respectively. As a result of this analysis, Admin PL will be included as a covariate in the model described in the next section.

Figure 4

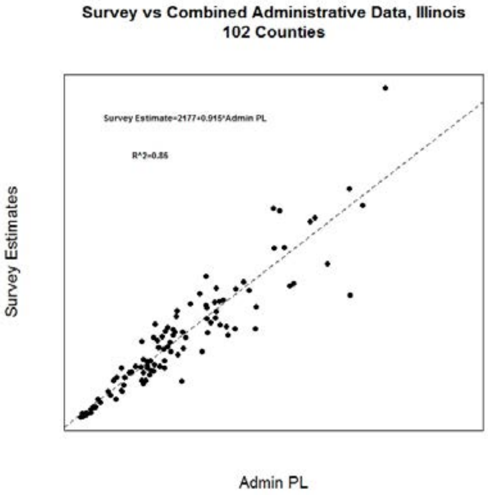


Table 2: Nationwide Summaries

	FSA			RMA			CDL			**Admin**		
	1st Qu.	Median	3rd Qu.	1st Qu.	Median	3rd Qu.	1st Qu.	Median	3rd Qu.	1st Qu.	Median	3rd Qu.
R^2	0.82	0.89	0.92	0.76	0.86	0.91	0.85	0.90	0.93	0.85	0.90	0.93
$\hat{b}$	0.85	0.91	0.99	0.89	0.97	1.17	0.75	0.84	0.91	0.75	0.84	0.89

3. Modeling Strategies

Following Erciulescu et al. (2018), the proposed model for a given state is a subarea-level model, where the area represents the ASD, the subarea represents the county and the subarea-level survey variances are treated as fixed and known. Of interest is prediction of planted acreage at the county and ASD levels.

3.1 Hierarchical Bayes Model

Let $i = 1, ..., m$ be an index for the m ASDs in the state under consideration; $j = 1, ..., n_i^c$, be an index for the n_i^c counties in ASD i; and n_{ij} be the sample size of the j^{th} county in the i^{th} ASD. The total number of counties in the state is $\sum_{i=1}^m n_i^c = n^c$ and the state sample size is $\sum_{i=1}^m \sum_{j=1}^{n_i^c} n_{ij} = n$.

Let $\hat{\theta}_{ij}$ be the survey estimate for county i in ASD j and $\hat{\sigma}^2_{ij}$ be the corresponding estimated survey variance. Illustrated for one state, one commodity and one parameter, the hierarchical Bayes subarea-level model is

$$\hat{\theta}_{ij}|(\theta_{ij}, \hat{\sigma}^2_{ij}, v_i) \overset{ind}{\sim} N(\theta_{ij}, \hat{\sigma}^2_{ij}), \tag{3}$$

$$\begin{aligned} \theta_{ij}|(v_i, \boldsymbol{\beta}, \sigma^2_u) &\overset{ind}{\sim} N(\mathbf{x}'_{ij}\boldsymbol{\beta} + v_i, \sigma^2_u), \\ v_i|\sigma^2_v &\overset{ind}{\sim} N(0, \sigma^2_v). \end{aligned} \tag{4}$$

To complete the Bayesian model specification, we consider a priori independent parameters and adopt noninformative, proper priors for $(\boldsymbol{\beta}, \sigma^2_u, \sigma^2_v)$. The least squares estimates of $\boldsymbol{\beta}$ are obtained from fitting a simple linear model for the county-level survey estimates against the county-level auxiliary information, and then used as parameters in the prior distribution for $\boldsymbol{\beta}$. In particular, we adopt a multivariate normal prior distribution for $\boldsymbol{\beta}$, with mean and variance denoted by the least squares estimate for the mean and the least squares estimate for the variance, multiplied by 10^3, respectively. By assigning a large prior variance, we adopt a diffuse prior for $\boldsymbol{\beta}$. The prior distributions for the model variance components σ^2_u and σ^2_v are $Uniform(0, 10^8)$ and $Uniform(0, 10^8)$, respectively. For more details on the prior distribution for the random-effects variance component see Browne and Draper (2005).

The model (3, 4) borrows information from all the counties in an ASD and from all the ASDs in the state, while combining auxiliary information available at the subarea level, $\mathbf{x}_{ij}$. The result model predictions are composite predictions, denoted by the weighted average of the subarea survey estimate and the best fitted values, after accounting for the area effect. That is, for a county j, in district i, the posterior mean is

$$\begin{aligned} \tilde{\theta}_{ij} &= \mathbf{x}'_{ij}\tilde{\boldsymbol{\beta}} + \tilde{\gamma}_i(\bar{\hat{\theta}}^{\gamma}_i - \bar{\mathbf{x}}^{\gamma'}_i\tilde{\boldsymbol{\beta}}) + \tilde{\gamma}_{ij}\left\{\hat{\theta}_{ij} - \mathbf{x}'_{ij}\tilde{\boldsymbol{\beta}} - \tilde{\gamma}_i(\bar{\hat{\theta}}^{\gamma}_i - \bar{\mathbf{x}}^{\gamma'}_i\tilde{\boldsymbol{\beta}})\right\} \\ &= \tilde{\gamma}_{ij}\hat{\theta}_{ij} + (1 - \tilde{\gamma}_{ij})\left\{\mathbf{x}'_{ij}\tilde{\boldsymbol{\beta}} + \tilde{\gamma}_i(\bar{\hat{\theta}}^{\gamma}_i - \bar{\mathbf{x}}^{\gamma'}_i\tilde{\boldsymbol{\beta}})\right\}, \end{aligned} \tag{5}$$

where $\tilde{\gamma}_{ij} = \frac{\tilde{\sigma}^2_u}{\tilde{\sigma}^2_u + \hat{\sigma}^2_{ij}}$, $\tilde{\gamma}_{i.} = \sum_{j=1}^{n^c_i} \tilde{\gamma}_{ij}$, $\tilde{\gamma}_i = \frac{\tilde{\sigma}^2_v}{\tilde{\sigma}^2_v + \tilde{\sigma}^2_u(\tilde{\gamma}_{i.})^{-1}}$, $\bar{\hat{\theta}}^{\gamma}_i = (\tilde{\gamma}_{i.})^{-1}\sum_{j=1}^{n^c_i} \tilde{\gamma}_{ij}\hat{\theta}_{ij}$ and $\bar{x}^{\gamma}_i = (\tilde{\gamma}_{i.})^{-1}\sum_{j=1}^{n^c_i} \tilde{\gamma}_{ij}x_{ij}$. Technical details on the derivation are provided in Appendix B.

A discussion on the choice of county-level covariate values x_{ij} is provided in the next subsection, as it depends on the availability of the data. When available, the county-level covariate values, x_{ij}, are Admin PL values constructed as described above, and the model is denoted by M. For comparison, a model with no covariates and a model with Admin PL constructed using only the FSA and the RMA data are also fit, and denoted by M0 and M1, respectively. In addition, the comparison of models M and M1 may be of interest to the agency because the current NASS process of setting official statistics uses FSA and RMA data, but it does not use CDL data directly; see Cruze et al. (2016) for a detailed description of the process.

3.2 Incomplete Data

Complete sets of data are needed to define the counties with corn planted acreage activity and for models (3, 4) to be fit. One other challenge in combining data from multiple sources is the incomplete availability of the data. For this, we develop modeling strategies to account for three cases of available information for a given county j, in ASD i:

1. $(\hat{\theta}_{ij}, \hat{\sigma}^2_{ij})$ are available, but x_{ij} is missing,

2. $(\hat{\theta}_{ij}, \hat{\sigma}^2_{ij}, x_{ij})$ are available,

3. $(\hat{\theta}_{ij}, \hat{\sigma}^2_{ij})$ are missing, but x_{ij} is available.

For the missing data cases, we assume the missing at random mechanism, since most of the missing covariate values correspond to smaller survey planted acreage estimates.

The first step in the modeling strategies is to impute the missing covariate values x_{ij}, for counties j in districts i, where survey estimates $(\hat{\theta}_{ij}, \hat{\sigma}^2_{ij})$ are available. For this, we use the x_{ij} values available for the most similar counties in the state. Similarity is defined using the absolute-value norm applied to the available survey estimates,

$$x_{ij} \leftarrow x_{ij'} \mid j^{'} = arg\ min_k \left\{ |\hat{\theta}_{ik} - \hat{\theta}_{ij}| \right\},$$

over all counties k with survey and auxiliary data available. The resulting set of counties n^c with survey and auxiliary data $(\hat{\theta}_{ij}, \hat{\sigma}^2_{ij}, x_{ij})$ available denotes all the counties with corn planting activity for the study.

After imputation, the models are fit to the n^c counties for which $(\hat{\theta}_{ij}, \hat{\sigma}^2_{ij}, x_{ij})$ are available, using R JAGS, and posterior distributions are constructed using MCMC simulation. We use 3 chains, each of 10,000 Monte Carlo samples, 1,000 burn-in samples and thinned every 9 samples. Convergence diagnostics are conducted for selected states. The convergence is monitored using trace plots, the multiple potential scale reduction factors (values less than 1.1) and the Geweke test of stationarity for each chain (Gelman and Rubin, 1992 and Geweke, 1992). Also, once the simulated chains have mixed, we construct the effective number of independent simulation draws to monitor simulation accuracy.

Using the chains of iterates obtained from the model fit, we construct posterior summaries from the posterior distributions of the nuisance parameters and of the county-level and ASD-level parameters of interest,

- nuisance parameters iterates : $\boldsymbol{\beta}^r, (\sigma^2_u)^r, (\sigma^2_v)^r$,
- county-level iterates: θ^r_{ij},
- district-level iterates: $\theta^r_i := \sum_{j=1}^{n^c_i} \theta^r_{ij}$,

where $r = 1, ..., R$, and R denotes the total MCMC iterates, after burn-in and thinning, equal to 3000 in the application study.

In the last step in the modeling strategies, the model output from the complete data fit is used to predict for counties where $(\hat{\theta}_{ij}, \hat{\sigma}^2_{ij})$ is missing but x_{ij} is available. For this, $\left\{\theta^r_{ij}\right\}_{r=1,...,R}$ are drawn from the linking model (4),

$$\theta^r_{ij} | (v^r_i, \boldsymbol{\beta}^r, (\sigma^2_u)^r) \overset{ind}{\sim} N(\mathbf{x}^{'}_{ij}\boldsymbol{\beta}^r + v^r_i, (\sigma^2_u)^r).$$

3.3 Consistency among Nested Levels

As discussed in the Section 1, NASS publishes the state-level value of corn planted acreage before estimation is conducted at the substate levels. To overcome the challenge of attaining consistency among predictions constructed for nested levels, we consider an external benchmarking adjustment, that is timely and practically usable. A detailed discussion of classic benchmarking adjustments is given in Rao and Molina (2015). Studies on different benchmarking adjustments to crop acreage prediction are discussed in Erciulescu et al. (2018). In this section, we illustrate a benchmarking adjustment applied to the model predictions constructed under the different data availability cases, so that the county-level predictions aggregate to the ASD-level predictions and the ASD-level predictions aggregate to the prepublished state-level value.

Raking provides a suitable benchmarking adjustment to ensure consistency of substate predictions with state targets. For this study, we use the extension of the classic ratio adjustment given in Erciulescu et al. (2018), and we apply the constraint at the (MCMC) iteration level. This type of benchmarking adjustment is not adopted as part of the prior information or the model, but it facilitates its application to the set of in-sample and not-in-sample counties, in a small amount of time.

Let the state-level target be denoted by a. Then the relation

$$\sum_{i,j}^{n^{c^*}} \tilde{\theta}_{ij}^{B} = a, \tag{6}$$

needs to be satisfied, where n^{c^*} is the total number of counties in the state and $\tilde{\theta}_{ij}^{B}$ is the final model prediction for county j and district i. Note that $n^{c^*} = n^c + (n^{c^*} - n^c)$, where n^c is the number of in-sample counties and $(n^{c^*} - n^c)$ is the number of not-in-sample counties. The ratio adjustment is applied at the MCMC iteration level as follows

$$\theta_{ij,r}^{B} := \theta_{ij,r} \times a \times \left(\sum_{k=1}^{m} \sum_{l=1}^{n_k^{c^*}} \theta_{kl,r} \right)^{-1}, \tag{7}$$

where $\theta_{ij,r}^{B}$ is the benchmarking-adjusted iteration, for $r = 1, ..., R$. Final county-level and district-level posterior summaries are constructed using the county-level iterates $\theta_{ij,r}^{B}$ and district-level iterates $\theta_{i,r}^{B} := \sum_{j=1}^{n_i^{c^*}} \theta_{ij,r}^{B}$. For example, the resulting posterior means/variances are constructed as Monte Carlo means/variances of iterates. The county-level and district-level posterior means satisfy the multi-level benchmaking to state-level target a; note that $n_i^{c^*}$ is the total number of counties in district i.

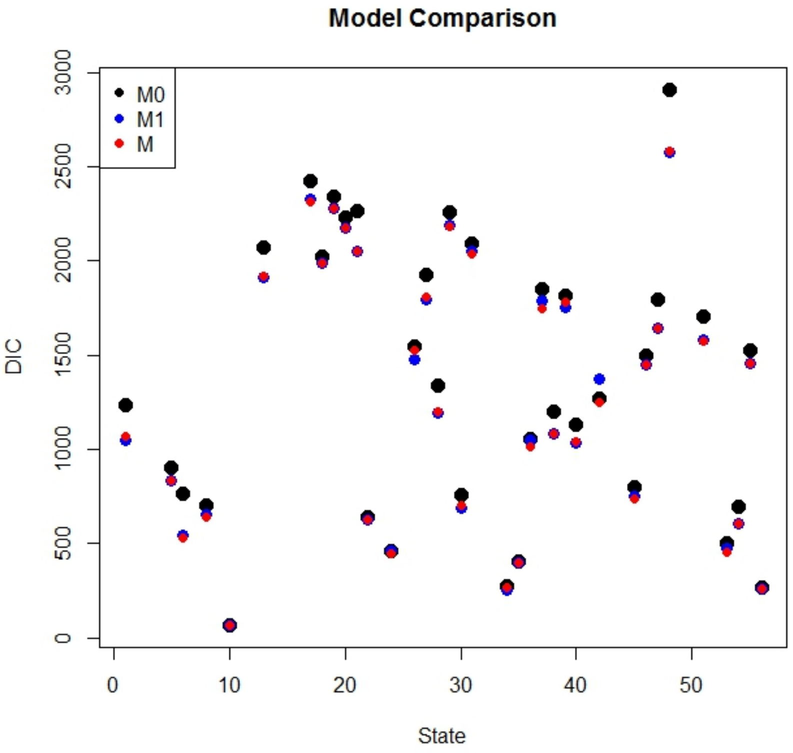
From (7), note the importance of correctly specifying the set of counties to be estimated, since a smaller/larger than the truth number of counties would result in an overadjustment/underadjustment in the predictions.

4. Results

Planted acreage data from the four sources summarized in Table 1 are used to define the set of counties to be estimated. For models fit and prediction, we define the set of counties with complete data after implementing the first step in the modeling strategies enumerated in Section 3.2. As previously mentioned, we consider three models for comparison: M0, the model fit to the survey data and no covariate; M1, the model fit to the survey data with covariate derived from FSA and RMA data (directly and imputed, when applicable); and M, the model fit to the survey data with covariate derived from FSA, RMA and CDL data (directly and imputed, when applicable). Note that the survey data modeled in all M0, M1 and M is the same, only the covariate data differ.

The goodness of fit for models M0, M1 and M, fitted state by state, is evaluated using the Deviance Information Criterion (DIC) and results are presented in Figure 5. The x-axis in Figure 5 illustrates the two-digit Federal Information Processing Standards (FIPS) codes for the 36 states, sampled for corn in 2015. Model comparison is conducted for each state, and not between states. The goodness of fit increases when auxiliary information is incorporated in the model, the best fit being when the Admin PL is defined using FSA, RMA and CDL. Models M1 and M result in similar performance; however, there are other benefits of using the CDL, as discussed in Section 4.2.

Figure 5



Models M0, M1 and M are further compared with respect to the contribution of auxiliary data to the final model predictions. Three-number summaries $(25\%, 50\%, 75\%$ quantiles) of the estimated shrinkage coefficients, $\tilde{\gamma}_{ij}$ (%) and $\tilde{\gamma}_i$ (%) defined for (5), are constructed over all the 36 states for which the models are fit and illustrated in Tables 3 and 4. Again, models M1 and M perform similarly. The auxiliary data and their relationship with the survey estimates receive larger weights in the final predictions under model M compared to model M0.

Increased Number of Reliable Estimates

Of great interest is the contribution of administrative data to increasing the number of county-level estimates. A nationwide map of the 2015 corn positive planted acreage county-level model predictions on the log10 scale, using model M, is illustrated in Figure 6. Model predictions are produced for 2627 counties, of which 2420 are in-sample counties

Table 3: Summary of Estimated Shrinkage Coefficients $\tilde{\gamma}_{ij}$ (%)

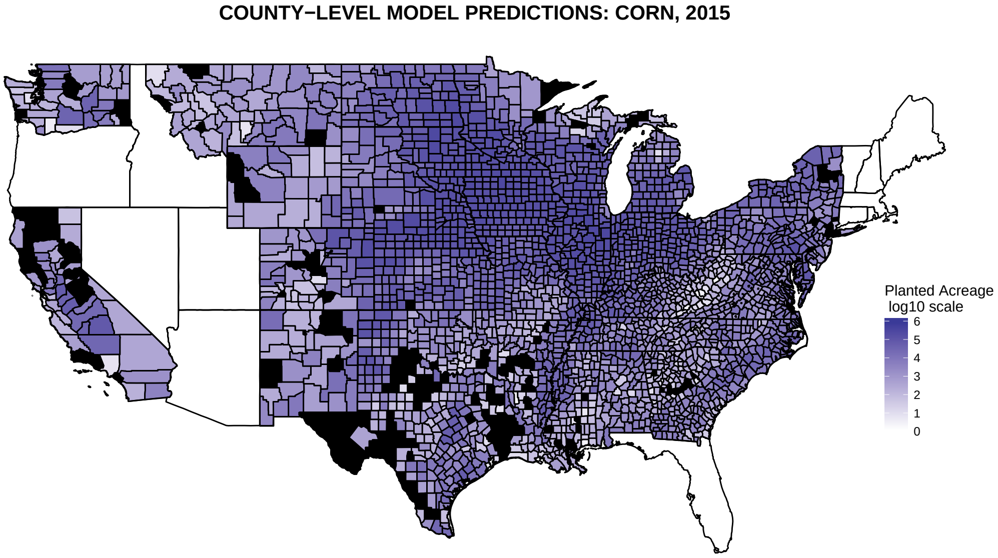
Approach	Covariate ADMIN PL	1st Qu.	Median	3rd Qu.
Model M0	None	60.66	85.69	98.01
Model M1	FSA and RMA	2.67	11.41	44.92
Model M	**FSA, RMA and CDL**	2.42	10.25	40.94

Table 4: Summary of Estimated Shrinkage Coefficients $\tilde{\gamma}_{i}$ (%)

Approach	Covariate ADMIN PL	1st Qu.	Median	3rd Qu.
Model M0	None	85.37	92.25	95.48
Model M1	FSA and RMA	46.04	62.13	77.36
Model M	**FSA, RMA and CDL**	47.90	66.35	82.54

and 207 are not-in-sample counties. Additionally, 121 model predictions were set to zero. Darker areas correspond to higher intensity regions. Not-in-sample predictions are mostly produced for counties located in non-major corn producing states and with small acreage amounts (the maximum not-in-sample model prediction is approximately 60% the median of the in-sample model predictions) and large CVs. In contrast, recall that survey estimates are available for 2426 counties, as illustrated in Figure 1.

Figure 6



Model Efficiency

Model efficiency comparisons are conducted for the set of counties where both a survey estimate and a model prediction are available. Compared to the survey estimates, the SEs and CVs of the model predictions are lower for most counties and districts. In Figure 7 we illustrate the reduction in CVs for the 2015 county-level estimates of corn planted acreage in Illinois, under model M.

In Tables 5 and 7 we illustrate nationwide results ($25\%, 50\%, 75\%$ quantiles), comparing the county-level survey SEs/CVs to the model SEs/CVs for models M1 and M. In Tables 6 and 8 we illustrate nationwide results ($25\%, 50\%, 75\%$ quantiles), comparing the district-level survey SEs/CVs to the model SEs/CVs for models M1 and M. Comparing a model's performance versus survey's performance based on precision/relative precision, we observe an increase in precision/relative precision in the range $34 - 70\%$/$32 - 72\%$ in most of the county-level SE/CV and in the range $27-57\%$/$48-54\%$ in most of the district-level SE/CV, with slight improvement at the county level for model M versus model M1. We do not see an overall increase in precision at the district level for model M versus model M1 because the districts are composed of both in-sample and not-in-sample counties, and more predictions for not-in-sample counties are constructed using model M1.

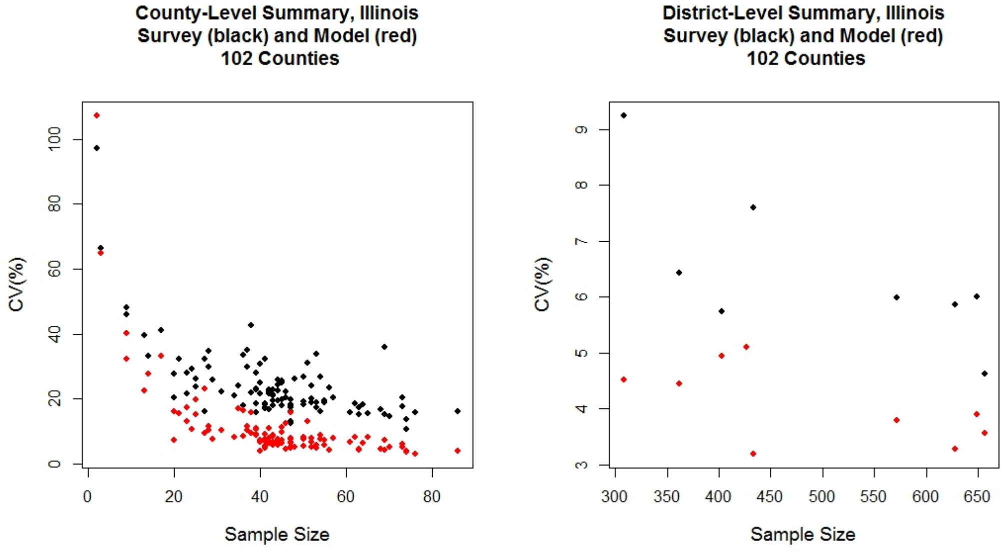


Table 5: SE Summaries for Counties with Available Survey Estimates

Approach	Covariate ADMIN PL	1st Qu.	Median	3rd Qu.
Survey		640.90	2719.00	9494.00
Model M1	FSA, RMA	429.40	1233.00	2850.00
Model M	**FSA, RMA and CDL**	429.30	1166.00	2839.00

Table 6: SE Summaries for Districts

Approach	Covariate ADMIN PL	1st Qu.	Median	3rd Qu.
Survey		4681.00	12220.00	36400.00
Model M1	FSA, RMA	2597.00	6121.00	15200.00
Model M	**FSA, RMA and CDL**	2958.00	6470.00	15310.00

The three-number summaries in Tables 5 - 8 do not reflect the relative efficiency at the domain (county or district) level. So, we report additional results in Figure 8, in the first row for 2420 counties with positive survey estimates and model predictions, and in the second row for the corresponding 272 districts (which may include additional model predictions); domains with relative efficiency values greater than 3 are removed. The relative SE/CV

Table 7: CV(%) Summaries for Counties with Available Survey Estimates

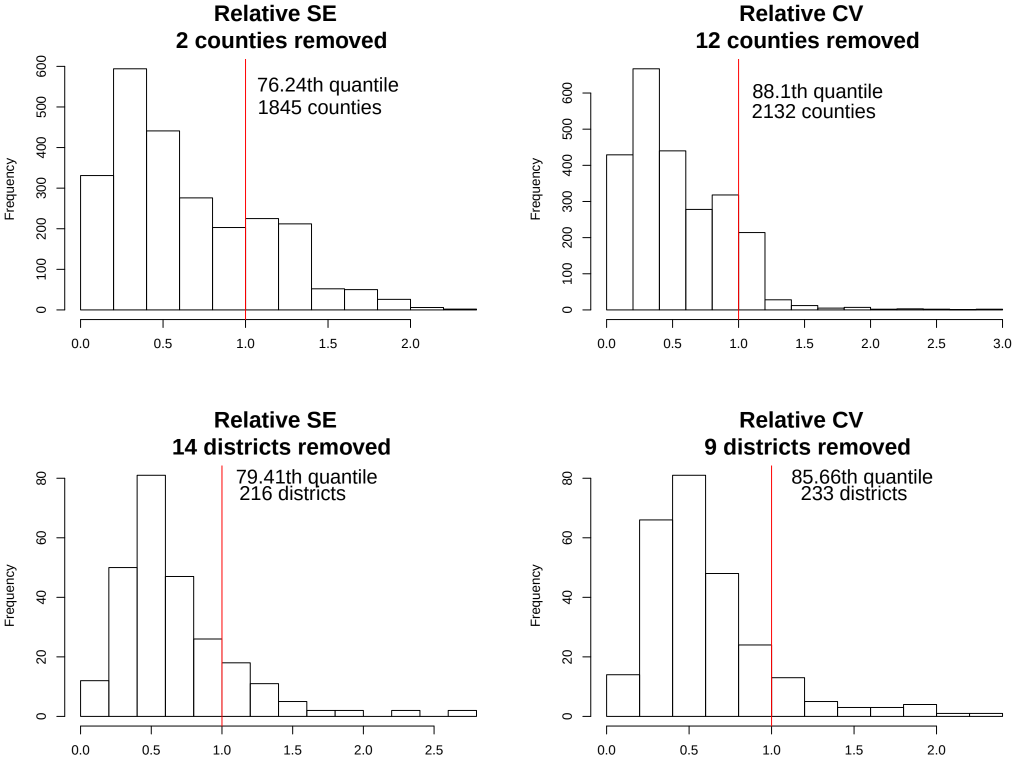
Approach	Covariate ADMIN PL	1st Qu.	Median	3rd Qu.
Survey		21.08	31.91	55.42
Model M1	FSA, RMA	5.97	12.60	38.74
Model M	**FSA, RMA and CDL**	5.90	11.84	37.92

Table 8: CV(%) Summaries for Districts

Approach	Covariate ADMIN PL	1st Qu.	Median	3rd Qu.
Survey		7.03	10.50	16.04
Model M1	FSA, RMA	3.19	4.58	8.19
Model M	**FSA, RMA and CDL**	3.22	4.73	8.50

is the ratio of the model prediction standard error/coefficient of variation to the survey estimate standard error/coefficient of variation. Values larger than one for the county-level relative SE are due to the benchmarking adjustments and values larger than one for the district-level relative SE are due to the not-in-sample predictions and to the benchmarking adjustments.

Figure 8


Relative SE
2 counties removed

Relative CV
12 counties removed

Relative SE
14 districts removed

Relative CV
9 districts removed

4.1 Official Statistics

The current NASS publication standard is based on the survey summary and on relative properties of the final estimates (the official statistics determined by NASS), for acreage and production; see page 117 in National Academies of Sciences, Engineering, and Medicine (2017) for more details. First, harvested acreage is defined as the difference between planted acreage and failed acreage, and yield is defined as the ratio of production to harvested acreage. A county is candidate for publication if a minimum of 30 valid positive reports of production or yield (respondents may report either quantity on the questionnaire) is available in the sample. If this sample size threshold is not met, then a county is a candidate for publication if the ratio between the sum of unweighted harvested acreage reports and the final harvested acreage estimate is greater than or equal to 25%, and based on at least 3 positive yield reports. Estimates are published or suppressed for all parameters (acreage, production and yield) with respect to each commodity in the county. Verifying nondisclosure limitations, the same publication standard may be applied to groups of counties, too. In 2015, NASS published estimates of corn for 1433 counties, which are available in NASS QuickStats (USDA NASS 2016b).

In (5), we provided the closed-form expression for the model predictions. Since they are composite predictions of various sources, the nationwide set of model predictions is candidate for official publication. However, the challenge in constructing fit-for-use official statistics is the need for a publication standard that would permit publication of model predictions. While the current publication standard may be adopted for the model predictions, it would not make use of other properties of the model predictions, such as standard errors or credible intervals. For this application study, we investigate a hypothetical CV-based assessment, consistent with the publication standards at other government agencies. For example, Marker (2016) reports that many government agencies use a CV-based assessment to determine the areas for which estimates can be published, with cutoffs typically in the range of $30\% - 50\%$. At NASS, Bell and Barboza (2012) consider an evaluation of CV-based publication standard and conclude that CVs should play a role in the publication of official statistics. We use a 30% threshold for the county-level CVs across the nation, leading to 1694 candidate county-level planted acreage predictions for publication of corn in 2015.

4.2 Discussion

In this paper the methodology developed was illustrated using corn planted acreage, and the results for 2015 were presented. As an external validation exercise, models with specification M1 were fit to data from years 2014, 2015 and 2016, on corn, soybean and sorghum, to each year-commodity-state combination. The number of model predictions is larger than the number of survey estimates in each case, but the number of predictions differs. For the counties and districts where both a model prediction and a official value were available, we compared the two using metrics such as median absolute difference, median absolute relative difference and credible interval coverage. In general, results indicated close agreement between the model predictions and the official values (constructed under the current NASS process).

As a consequence of the model specification, in particular the normality assumption in the linkage model, predictions are set to zero in some counties because the posterior means were negative. While we acknowledge that other choices of distributions may be considered, for example lognormal, we recognize the simplicity of the current specification, especially with respect to prediction and benchmarking at multiple levels of interest. Also, the zero value was a reasonable choice for the counties where posterior means were negative, as corresponding to posterior quantiles close to 50%.

Among the states sampled for corn, soybean, sorghum, winter wheat in 2015, the largest numbers of not-in-sample predictions are for Texas, Texas, Mississippi, Georgia, where, respectively, 42, 70, 28 and 38 out of 184, 122, 73 and 154 counties were predicted. Their total planted acreage accounted for approximately 0.7$\%$, 11.83$\%$, 5.23$\%$ and 12.47$\%$, respectively, of total planted acreage in the state. See Appendix D for additional results on soybean, sorghum and winter wheat. Hence, benchmarking only the set of counties where survey estimates are available would have resulted in overadjusting the predictions. While the proportion of total acreage accounted for by the not-in-sample counties is small, the predictions play an important role in setting predictions for other variables of interest, such as harvested acreage, production and yield.

5. Conclusions

The quality and usability of different data sources are discussed, and contributions of administrative data are illustrated for the 2015 corn planted acreage estimation study. Some results are presented for additional commodities: soybean, sorghum and winter wheat. Blending survey and administrative data, we produce model county-level and district-level predictions for a set of counties predefined using in-sample data available from the survey summary and not-in-sample data available from administrative sources.

Our initial contribution is to use the administrative data to determine the set of subareas with crop-specific planting activity in 2015. This approach is novel and we encourage similar investigations for other small area estimation applications where small domain characteristics are diverse within the large domains and not-in-sample predictions are of interest, such as agricultural applications (i.e. county-level cash rental rate estimation makes sense only for counties where at least one cash rental contract exists), health applications (i.e. youth smoking prevalence estimation make sense only for domains where at least one youth smoker actually exists) or education applications (i.e. estimation of American Indian children ageing 5-17 in poverty makes sense only for domains where at least one American Indian child ageing 5-17 lives).

For the methodology illustrated, we presented the implicit subarea-level weights associated with the different components of the final prediction. The contribution of administrative data to final predictions was evaluated using the shrinkage parameter γ_{ij}. Model specifications, using a covariate derived from FSA and RMA data alone (M1), or from FSA, RMA and CDL data (M) are compared. Model M is slightly more efficient than model M1; however, it is important to note that, under model M1, 110 county-level Admin PL values were imputed, while under model M, only 11 county-level Admin PL values were imputed. Alternative strategies for imputation of missing auxiliary values and accounting for different errors in the covariates in the final model predictions are of interest for future research.

The number of positive model predictions is larger than the number of available survey estimates; however, a more robust model specification is of future interest, to avoid the construction of negative predictions. Under model specification M, a total of 2629 county-level predictions were constructed, while under model M1, only 2486 county-level predictions were constructed. For the set of counties where survey estimates are available and positive model predictions are constructed, the county-level and district-level model SEs and CVs are lower than the corresponding survey SEs and CVs, respectively.

Another major contribution of this paper is the operational framework presented, as it applies to any small area estimation application, from data preparation and challenges in dealing with specific features and incompleteness, to constructing a pool of predictions as candidates for official statistics and challenges associated with the publication process. Areas of improvement include exploration of state-specific, commodity-specific and time-specific covariates, revision of the set of counties with indicated corn planted acreage activity and improvement of the prediction by accounting for key periods of time in the crop development.

Finally, the current NASS publication standard is being revised. Literature review indicates CV-based publication standards for other government agencies around the world. In this paper, we investigated the effect of a simple threshold of 30% on the model CV, but concurrent research is being conducted; see Cruze et al. (2018).

6. Disclaimer and Acknowledgements

The Findings and Conclusions in This Preliminary Publication Have Not Been Formally Disseminated by the U. S. Department of Agriculture and Should Not Be Construed to Represent Any Agency 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.

The authors are grateful to Habtamu Benecha, Valbona Bejleri, Arthur Rosales, Claire Boryan, Rick Mueller, and Linda Young from the National Agriculture Statistics Service and to Nell Sedransk from the National Institute of Statistical Sciences, for useful discussions or comments.

References

Bell J., and Barboza W. (2012), “Evaluation of Using CVs as a Publication Standard.” Paper presented at the Fourth International Conference on Establishment Surveys, Montreal, Quebec, Canada, June 11-14.

Boryan, C., Yang Z., Mueller, 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.

Cruze N.B., Erciulescu A.L., Nandram B., Barboza W.J., Young L.J. (2016), “Developments in Model-Based Estimation of County-Level Agricultural Estimates.” *International Conference on Establishment Surveys V Proceedings. Alexandria, VA: American Statistical Association*. Available at *http://ww2.amstat.org/meetings/ices/2016/proceedings/131_ices15Final00229.pdf*.

Cruze N.B., Erciulescu A.L., Benecha H., Bejleri V., Nandram B., Young L.J. (2018), “Toward an Updated Publication Standard for Official County-Level Crop Estimates.” *Joint Statistical Meetings Proceedings. Government Statistics Section. Alexandria, VA: American Statistical Association.* To appear.

Erciulescu A.L., Cruze N.B., Nandram B. (2016), "Model-Based County-Level Crop Estimates Incorporating Auxiliary Sources of Information." *Joint Statistical Meetings Proceedings. Survey Research Methods Section. Alexandria, VA: American Statistical Association,* 3591-3605.

Erciulescu A.L., Cruze N.B., Nandram B. (2017), "Small Area Estimates for End-Of-Season Agricultural Quantities." *Joint Statistical Meetings Proceedings. Survey Research Methods Section. Alexandria, VA: American Statistical Association,* 541-560.

Erciulescu A.L., Cruze N.B., Nandram B. (2018) "Model-Based County-Level Crop Estimates Incorporating Auxiliary Sources of Information, " *Journal of the Royal Statistical Society, Series A*, DOI 10/1111/rssa.12390.

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

Gelman, A. and Rubin, D.B. (1992) "Inference from iterative simulation using multiple sequences," *Statistical Science*, 7, 457-511.

Geweke, J. (1992) "Evaluating the accuracy of sampling-based approaches to calculating posterior moments," *Bayesian Statistics 4 (ed JM Bernado, JO Berger, AP Dawid and AFM Smith). Clarendon Press, Oxford, UK.*

Kennedy C., Mercer A., Keeter S., Hatley N., McGeeney K., and Gimenez A. (2016), "Evaluating Online Nonprobability Surveys," *Pew Research Center*, available at *http://www.pewresearch.org/2016/05/02/evaluating-online-nonprobability-surveys/.*

Kim J.K., Wang Z., Zhu Z., Cruze N.B. (2018), "Combining Survey and Non-Survey Data for Improved Sub-Area Prediction Using a Multi-Level Model," *Journal of Agricultural, Biological, and Environmental Statistics*, 23, 2, 175-189.

Marker D. (2016), "Presentation to National Academy of Sciences Panel on Crop Estimates," *Unpublished presentation*. National Academy of Sciences report available at *https://www.nap.edu/catalog/24892/improving-crop-estimates-by-integrating-multiple-data-sources*.

National Academies of Sciences, Engineering, and Medicine (2017), "Improving Crop Estimates by Integrating Multiple Data Sources," *Washington, DC: The National Academies Press*, https://doi.org/10.17226/24892.

Parsons J. (1996), "Estimating the Coverage of Farm Service Agency Crop Acreage Totals," *USDA NASS Research Report*, SRB-96-02.

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 sub-area level model," *Journal of Multivariate Analysis*, 127, 36-55.

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

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

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

USDA NASS APS (2018), "Crops/Stocks Agricultural Survey," *https://www.nass.usda.gov/Surveys/Guide_to_NASS_Surveys/Crops_Stocks/index.php*.

USDA NASS CAPS (2018), "County Agricultural Production," *https://www.nass.usda.gov/Surveys/Guide_to_NASS_Surveys/County_Agricultural_Production/index.php*.

USDA NASS CDL (2018), "CropScape and Cropland Data Layers - FAQs," *https://www.nass.usda.gov/Research_and_Science/Cropland/sarsfaqs2.phpSection1_9.0*

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

Appendix A. Posterior Mean Derivation

It is easy to show that

$$\theta_{ij}|(v_i, \boldsymbol{\beta}, \sigma_u^2, \hat{\theta}_{ij}, \hat{\sigma}_{ij}^2) \overset{ind}{\sim} N(\gamma_{ij}\hat{\theta}_{ij} + (1-\gamma_{ij})(\mathbf{x}_{ij}^{'}\boldsymbol{\beta} + v_i), (1-\gamma_{ij})\sigma_u^2)$$

and

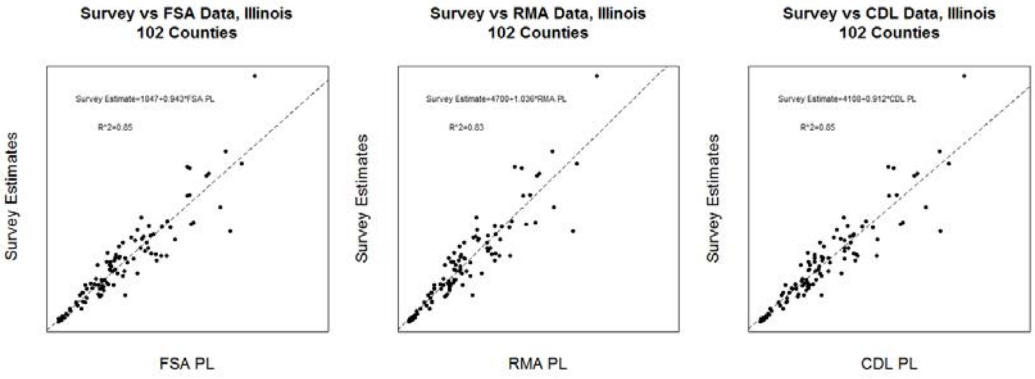
$$v_i|(\boldsymbol{\beta}, \sigma_u^2, \sigma_v^2, \hat{\theta}_{ij}, \hat{\sigma}_{ij}^2) \sim N(\gamma_i(\bar{y}_i^{\gamma} - \bar{\mathbf{x}}_i^{\gamma'}\boldsymbol{\beta}), (1-\gamma_i)\sigma_v^2),$$

where $\gamma_{ij} = \sigma_u^2(\hat{\sigma}_{ij}^2 + \sigma_u^2)^{-1}$ and $\gamma_i = \sigma_v^2(\sigma_v^2 + \sigma_u^2(\sum_{j=1}^{n_i^c}\gamma_{ij})^{-1})^{-1}$, for $i = 1, ..., m, j = 1...n_i^c$.

The result follows.

Appendix B. Borrow Information from Multiple Data Sources

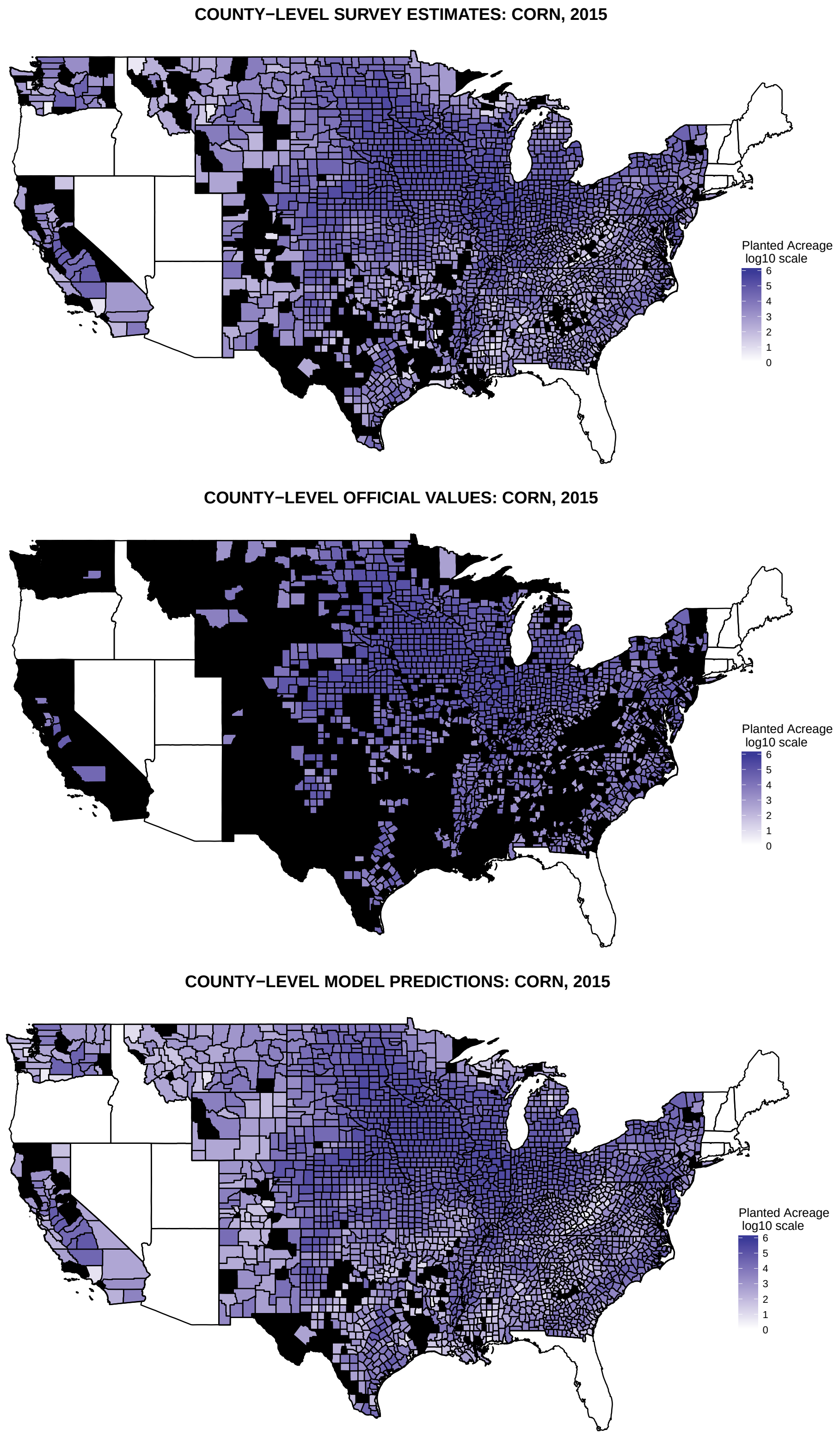
Figure 9



Appendix C. Increased Number of Reliable Estimates for Other Commodities

The county-level maps in Figures 10-13 depict positive survey (CAPS) estimates, official values and model (M) predictions on the log10 scale, for corn, soybean, sorghum and winter wheat, respectively. Dark areas correspond to high intensity regions.

Figure 10



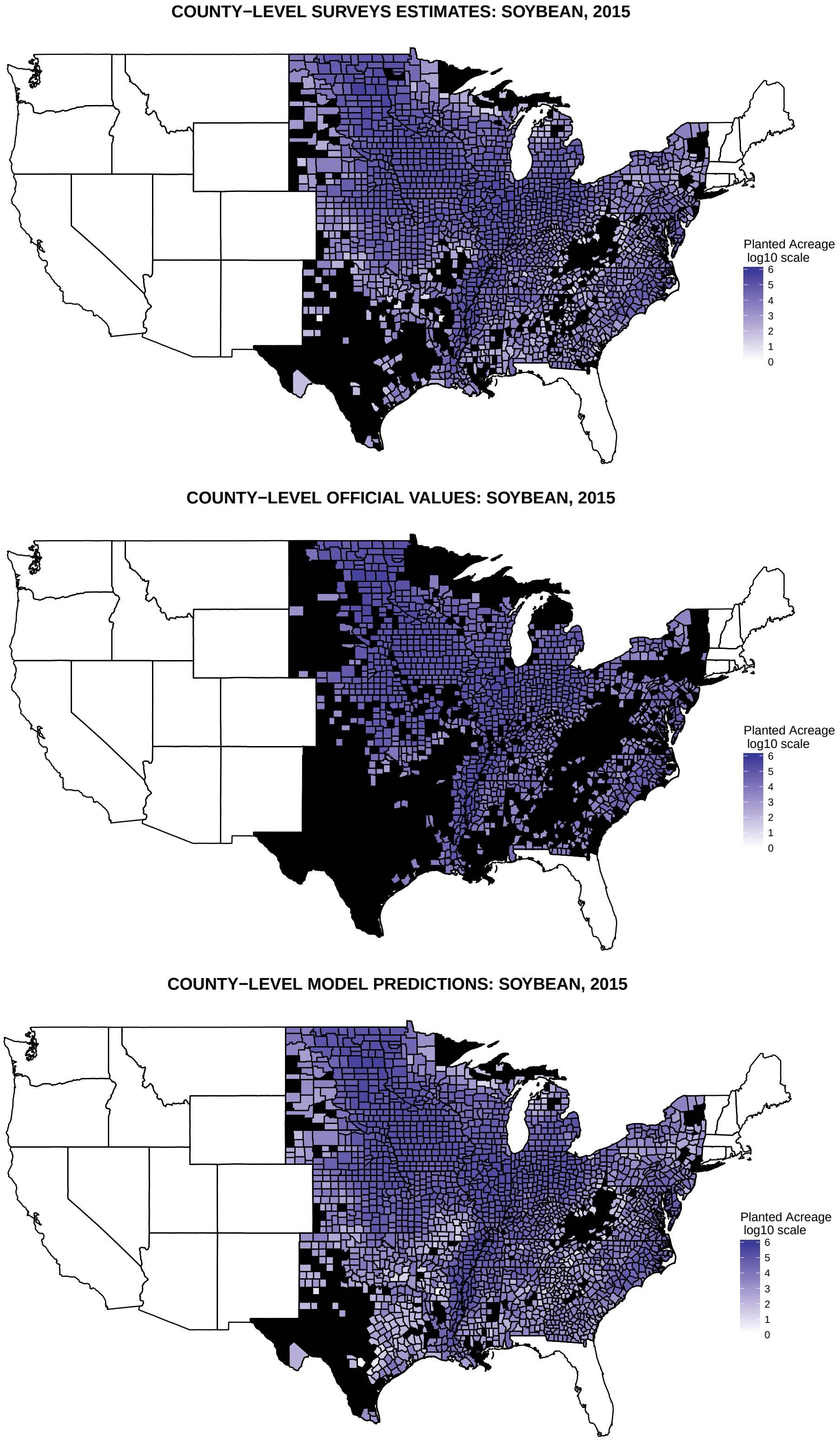
COUNTY−LEVEL SURVEY ESTIMATES: CORN, 2015

COUNTY−LEVEL OFFICIAL VALUES: CORN, 2015

COUNTY−LEVEL MODEL PREDICTIONS: CORN, 2015

- 1433 official values
- 2426 survey estimates; 1125 have CVs ≤ 30%
- 2627 model predictions; 1694 have CVs ≤ 30%
    - Texas: largest number of not-in-sample predictions, 42 out of 184 counties, accounting for ~0.7% of planted acreage in the state
    - 121 zero predictions

Figure 11



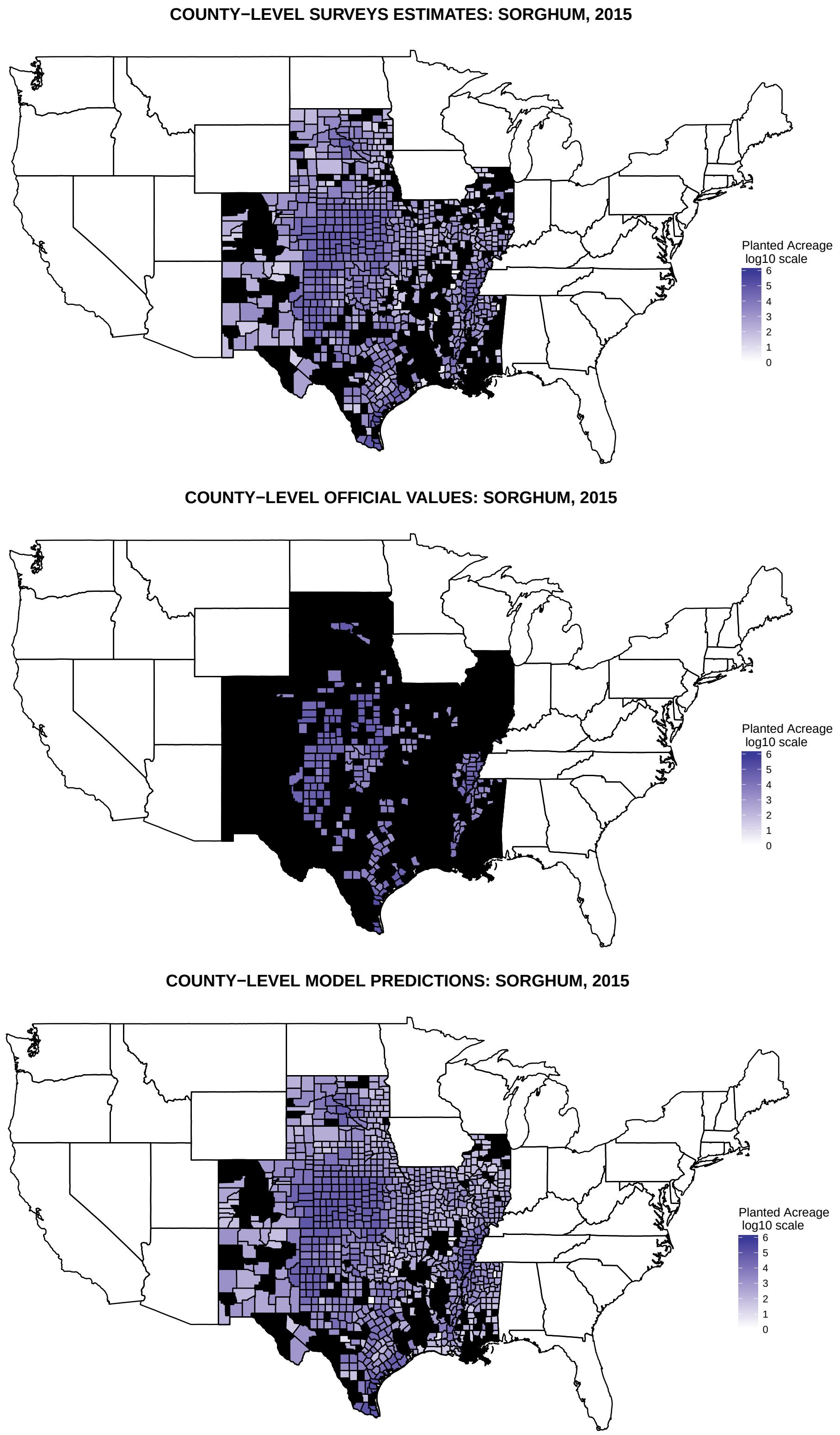
COUNTY−LEVEL SURVEYS ESTIMATES: SOYBEAN, 2015

COUNTY−LEVEL OFFICIAL VALUES: SOYBEAN, 2015

COUNTY−LEVEL MODEL PREDICTIONS: SOYBEAN, 2015

- 1306 official values
- 2012 survey estimates; 1046 have CVs ≤ 30%
- 2224 model predictions; 1472 have CVs ≤ 30%
    - Texas: largest number of not-in-sample predictions, 70 out of 122 counties, accounting for ~11.83% of planted acreage in the state
    - 173 zero predictions

Figure 12



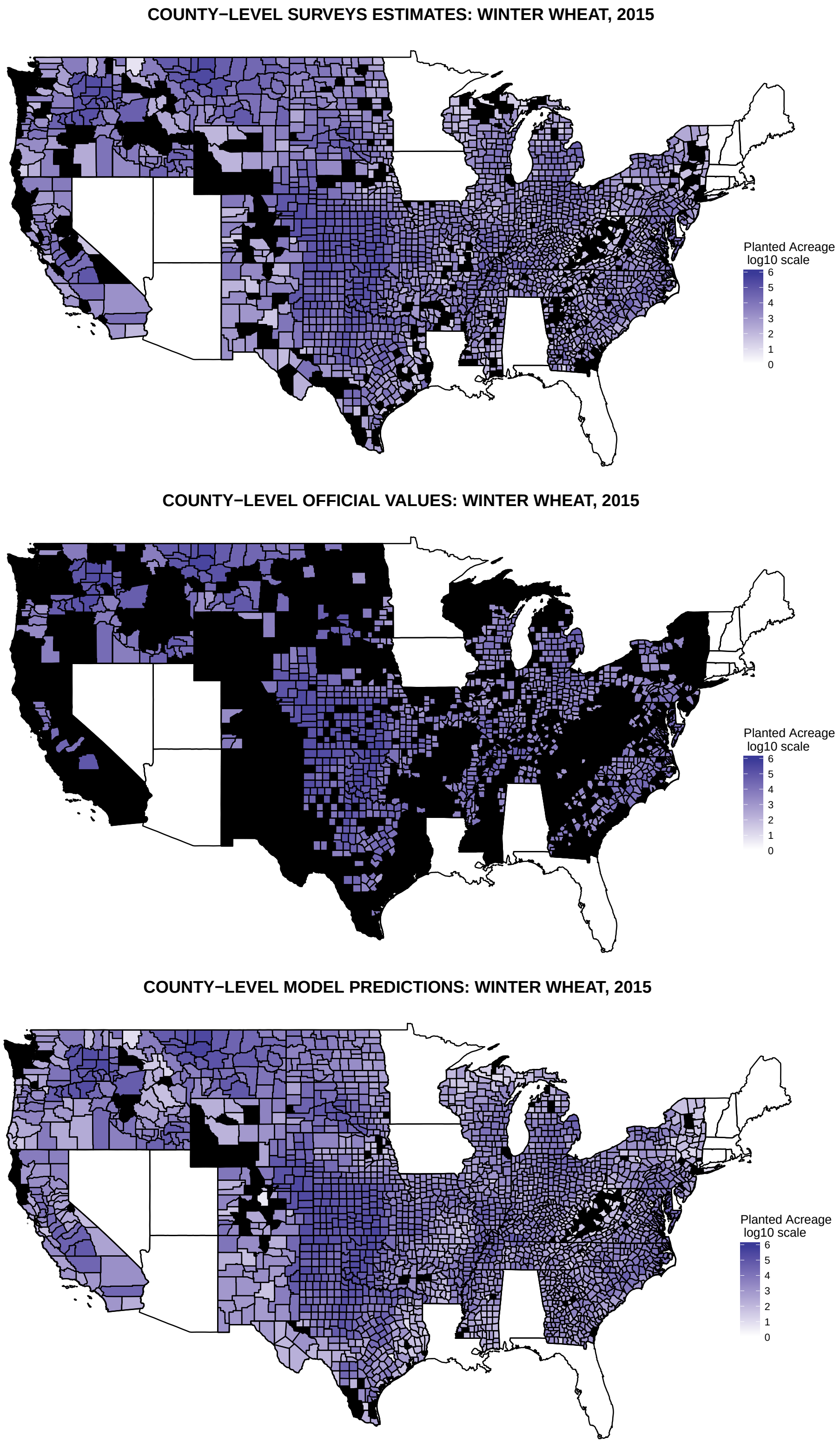
COUNTY−LEVEL SURVEYS ESTIMATES: SORGHUM, 2015

COUNTY−LEVEL OFFICIAL VALUES: SORGHUM, 2015

COUNTY−LEVEL MODEL PREDICTIONS: SORGHUM, 2015

- 218 official values
- 754 survey estimates; 135 have CVs ≤ 30%
- 922 model predictions; 390 have CVs ≤ 30%
  - Mississippi: largest number of not-in-sample predictions, 28 out of 73 counties, accounting for ~5.23% of planted acreage in the state
  - 89 zero predictions

Figure 13


COUNTY−LEVEL SURVEYS ESTIMATES: WINTER WHEAT, 2015

COUNTY−LEVEL OFFICIAL VALUES: WINTER WHEAT, 2015

COUNTY−LEVEL MODEL PREDICTIONS: WINTER WHEAT, 2015

- 1049 official values
- 2191 survey estimates; 697 have CVs ≤ 30%
- 2417 model predictions; 1321 have CVs ≤ 30%
  - Georgia: largest number of not-in-sample predictions, 38 out of 154 counties, accounting for ~12.47% of planted acreage in the state
  - 64 zero predictions