sciendo

Journal of Official Statistics, Vol. 36, No. 1, 2020, pp. 63–88, http://dx.doi.org/10.2478/JOS-2020-0004

Statistical Challenges in Combining Survey and Auxiliary Data to Produce Official Statistics

Andreea L. Erciulescu[1], *Nathan B. Cruze*[2], *and Balgobin Nandram*[3]

Combining survey and auxiliary data to produce official statistics is gaining interest at federal agencies and among policy makers due to its efficiency. Recent studies have shown the practicality of small area estimation modeling approaches in the context of integrating data from multiple sources to improve estimation at fine levels of aggregation. In this article, agricultural predictions are constructed using a hierarchical Bayes subarea-level model, fit to data available from different sources. Auxiliary data are initially used to complement the survey data and define the prediction space, and then to define covariates for the model. Finally, not-in-sample predictions are constructed using the model output, and benchmarking constraints are imposed on the final set of in-sample and not-in-sample predictions. Unlike most of the studies discussing not-in-sample prediction, this article illustrates a method that uses the data available from multiple sources to define the prediction space. As a consequence, the resulting framework provides a larger set of nationwide predictions as candidate for official statistics, and extrapolation is not of concern. Challenges in developing the methods to combine different data sources are discussed in the context of planted acreage prediction.

Key words: Administrative data; benchmarking; incomplete data; not-in-sample prediction; 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. In this article, we consider novel ways of using administrative data in the process of constructing official statistics. Specifically, administrative data are used to complement the survey data and define the set of domains for which predictions are needed. Then,

[1] Westat, 1600 Research Blvd., Rockville M.D., U.S.A. Email: alerciulescu@gmail.com
[2] USDA National Agricultural Statistics Service, Research and Development Division, 1400 Independence Avenue, SW, Washington D.C., U.S.A. Email: nathan.cruze@usda.gov
[3] Worcester Polytechnic Institute, Mathematical Sciences, Stratton Hall, 100 Institute Road, Worcester, MA 01609-2247, Massachusetts, 01609, U.S.A. Email: balnan@wpi.edu
Acknowledgments: 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 work of Dr. Erciulescu was conducted while she was a Research Associate with the National Institute of Statistical Sciences, working at the National Agriculture Statistics Service. The authors thank the Associate Editor and the two referees for their feedback and questions that led to an improved manuscript.

models that integrate survey and administrative data are used to construct predictions for domains with survey sample sizes as small as zero. This work builds on a series of research studies conducted at the United States Department of Agriculture's (USDA's) National Agricultural Statistics Service (NASS) to innovate the current methods of setting official statistics for acreage, production and yield at state and substate levels of aggregation. 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 district) is defined as a group of contiguous counties within a state.

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 ($k = 1, \ldots, m$) is an index for the small areas. The small area parameters of interest θ_k are estimated using a linking model,

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

where z_k are area-level covariates with p components, including an intercept, and (β, σ_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 (β, σ_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 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 ($j = 1, \ldots, n_i^c$) is an index for the subareas, i ($i = 1, \ldots, m$) 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 subarea mean θ_{ij}, which is estimated using a hierarchical linking model,

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

where x_{ij} are subarea-level covariates with p components, including an intercept, and $(\beta, \sigma_u^2, \sigma_v^2)$ 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 linking model in Torabi and Rao (2014) to allow for a hierarchical level for parameters β and to remove distributional assumptions in the first hierarchical level. Erciulescu et al. (2018, 2019) studied the FG model using a hierarchical Bayes framework, adopting prior distributions for $(\beta, \sigma_u^2, \sigma_v^2)$.

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, that is the estimates exist and are in the parameter space (positive total acreage estimates and positive sampling variances). However, for the not-in-sample subareas (domains with missing survey data), inference conducted relies on the linking model's specification. Given (1), a typical choice of estimator for the not-in-sample areas is the synthetic estimator $\mathbf{z}_k'\boldsymbol{\beta}$, 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 linking model (1) or (2), for area-level or subarea-level, respectively.

Building on the work of Erciulescu et al. (2019), we combine survey and auxiliary data and use a subarea-level model to construct planted acreage predictions for a set of counties defined by the union of all the available data sources. Statistical challenges and breakthroughs in combining data from multiple sources to produce official statistics are discussed throughout the paper. In particular, we identify a common geographic level and time point to combine data from a probability survey with nonprobability data from three administrative sources, the latter lacking uncertainty measures. As in Erciulescu et al. (2019), we treat the auxiliary data as fixed and free of error, but details on potential error sources in these data are available in Erciulescu et al. (2019). Note that Erciulescu et al. (2019) investigated these sources only for predictive power, and used only one at a time in developing the models (to avoid multicollinearity problems). Also, the authors tackled prediction for harvested acreage only for counties with both sample and administrative data available. In this article, we integrate all the data to identify the set of counties with planting activity for a specific crop (or the prediction space), in a given crop season, and to construct a covariate with good predictive power and observations available for all the counties in the prediction space. Challenges in multistage, nationwide prediction for counties with sample sizes as small as zero (not the case in Erciulescu et al. 2019) are addressed using hierarchical Bayes subarea-level models.

Modeling strategies are developed to deal with incomplete data and benchmarking methods are implemented to overcome the challenge of attaining consistency among predictions at nested levels of aggregation. Whereas Erciulescu et al. (2019) developed and compared models for direct estimates scaled by the sample sizes, with a hierarchy for sampling variances and different benchmarking methods, here we adopted the model for the direct estimates, with fixed sampling variances and the ratio adjustment, as a practical method with good performance that allows for prediction for subareas with sample sizes of just one or even zero where suggested by auxiliary information. This outcome was not possible under the model specification pursued in Erciulescu et al. (2019). Moreover, due to the extended prediction space, the possible over-adjustment due to benchmarking is less of a concern for NASS than it was for Erciulescu et al. (2019). As a result, a crop-specific framework of producing predictions is presented, with the potential to increase the number of official statistics constructed using current methodology.

In summary, the major contributions of this article are as follows:

- integration of all the available data to define the prediction space, as well as a covariate with good predictive power;

- modeling strategies to deal with incomplete administrative data and missing at random (MAR) assumption;
- not-in-sample prediction;
- reduction in over-adjustment due to benchmarking; and
- increase in the number of official statistics, given a common criterion.

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 a certain crop, that is, county-level corn, as in the case study illustrated here. In Section 3, modeling strategies addressing different scenarios of available data and the corresponding derived predictors are presented. 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. A discussion is provided in Section 5. Additional results on corn, soybean, sorghum and winter wheat are presented in the Appendix (Section 6).

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. Next, we introduce a method that combines survey data and administrative data to identify the 2015 in-sample and not-in-sample counties of interest for corn planted acreage prediction. Finally, we investigate the potential for using auxiliary data as covariates in hierarchical models. 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.

2.1. NASS Survey Data

The probability sample of interest in 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. 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 the 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.

County-level and district-level survey estimates and associated variance estimates are available from the NASS's CAPS summary. The district-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 district-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 2,837 counties, and NASS produced survey estimates for 2,426 in-sample counties. Survey estimates are not available for the remaining 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 with a black dot. The counties with zero planted acreage predictions and not-in-sample counties for corn in 2015 are represented in white. 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 areas correspond to high acreage intensity regions, in particular the Midwestern corn belt states.

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 is 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-district-to-state agreement to hold. More details on the benchmarking adjustment we utilize are presented in Subsection 3.3.

The number of counties and districts vary across the states and across commodities. For 2015 corn, the number of counties within districts ranges from 1 to 32, with a median of 8 and the number of districts 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 district 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

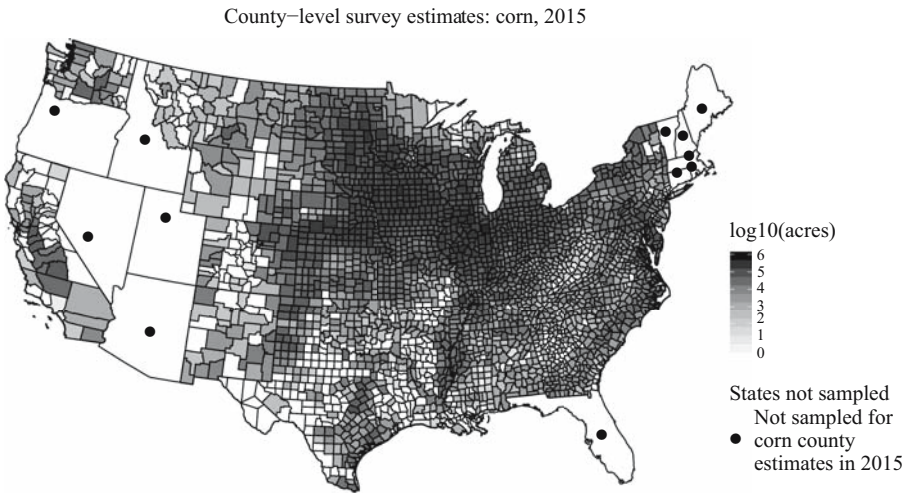


Fig. 1. Nationwide map of the end-of-season positive county-level planted acreage survey estimates available for corn in 2015 from the NASS CAPS summary, with all non-zero estimates on the log(10) scale.

positive acreage was reported. The county sample size differs from state to state and commodity to commodity. For 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 sample size ranges for Illinois are illustrated on the x-axes in Figure 2.

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 district-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.

2.2. *Auxiliary Data*

We explore auxiliary data, available from three USDA agencies: NASS, the Farm Service Agency (FSA) and the Risk Management Agency (RMA). FSA administers US farm programs, such as county-level revenue loss protections (USDA FSA 2019). RMA oversees the Federal Crop Insurance Corporation, which provides crop insurance to participating farmers and agricultural entities (USDA RMA 2019). For this, FSA and RMA collect data from farmers participating in such programs. NASS produces the Cropland Data Layer (USDA NASS CDL 2018), a crop-specific land cover product that uses satellite and FSA ground-reference data to classify crop types in 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

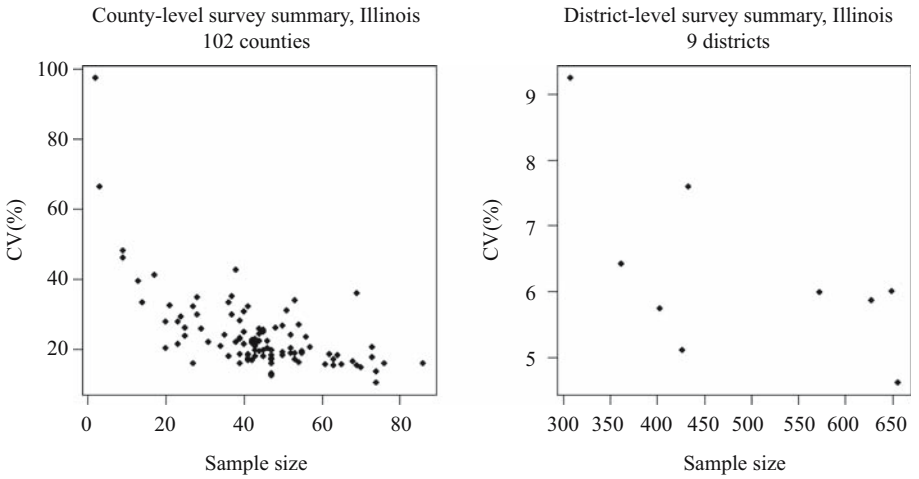


Fig. 2. Plots of CVs of the 2015 survey county-level and district-level estimates of planted acreage of corn in Illinois against corresponding sample sizes.

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, 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 and comprising the nonprobability sample data under consideration.

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 assume the nonprobability county-level values from FSA, RMA and CDL 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 2,726 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 2,748 counties. The number of counties with corn planting activity differs across years, states, commodities and data sources.

The county-level quantity of interest is the total planted acreage and the values available from the three sources (FSA, RMA, CDL) of auxiliary data are measurements of the same county-level quantity, that is corn total planted acreage. It is known that all three sources may suffer from downward biases (see Cruze et al. 2019 for a literature review of geography and remote sensing studies). As an attempt to avoid the possible downward bias and obtain a covariate with good predictive power for total county-level acreage, we combine the three sources to construct one set of values indicating the maximum number of available corn planted acreages, reported by volunteers or remotely classified. 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,

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

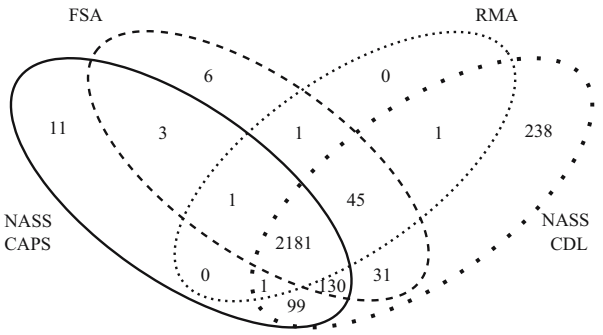


Fig. 3. Counties, in sampled states, with corn planting activity, 2015.

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

As expected, nationwide analysis indicates strong linear relationships between the survey estimates and the administrative data for all the 36 states. For each sampled state, a simple regression model was fit to the survey estimates, using intercept and FSA, RMA, CDL or Admin PL as predictor. Summaries of R^2 values and estimated slope coefficients $\hat{b}$, for all the states, are reported in Table 2 (25%, 50%, 75% quantiles).

In Figure 4 we display the linear fit between the survey estimates and the derived administrative values, Admin PL, and in Figure 5 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.

3. Modeling Strategies

The proposed model for a given state is a subarea-level model, where the area represents the district, 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 district levels. Prediction is conducted state by state and commodity by commodity, for all counties within states identified to have planted acreage activity in the given crop season.

3.1. Hierarchical Bayes Model

Let $i = 1, \ldots, m$ be an index for the m districts in the state under consideration; $j = 1, \ldots, n_i^c$, be an index for the n_i^c counties in district i; and n_{ij} be the sample size of the j^{th} county in the i^{th} district. 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 (total planted acreage) survey estimate for county i in district j and $\hat{\sigma}_{ij}^2$ be the corresponding estimated survey variance. For now, assume that county-level covariate

Table 2. Nationwide summaries for linear regression models applied to the data for every sampled state.

	FSA			RMA			CDL			Admin		
	1st Quantile	Median	3rd Quantile	1st Quantile	Median	3rd Quantile	1st Quantile	Median	3rd Quantile	1st Quantile	Median	3rd Quantile
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

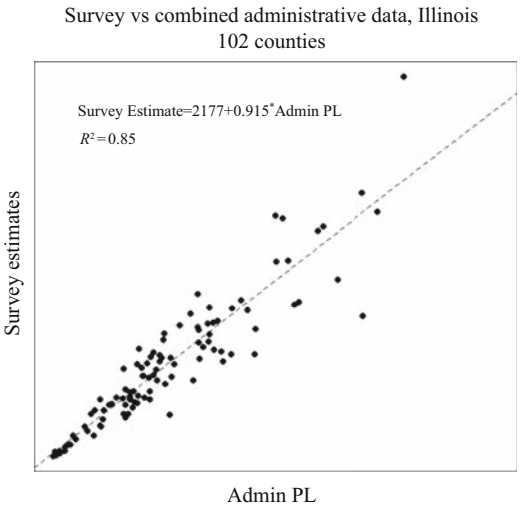


Fig. 4. Plot of survey estimates against derived administrative data values of planted acreage of corn (maximum value from available administrative sources) overlaid with best simple linear regression line.

values x_{ij} are available; a discussion on the availability of such covariates is provided later. Illustrated for one state, one commodity and one parameter, the hierarchical Bayes subarea-level model is

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

$$\theta_{ij}|(v_i, \beta, \sigma_u^2) \overset{ind}{\sim} N(\mathrm{x}_{ij}'\beta + v_i, \sigma_u^2), \tag{4}$$

$$v_i|\sigma_v^2 \overset{ind}{\sim} N(0, \sigma_v^2). \tag{5}$$

The parameters $(\beta, \sigma_u^2, \sigma_v^2)$ are assumed independent a priori, for which noninformative, proper priors are adopted. The least squares estimates of β are obtained from fitting a simple linear model for the county-level survey estimates against the county-level auxiliary information, and then used as fixed and known parameters in the prior distribution for β. In particular, we adopt a multivariate normal prior distribution for β, 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

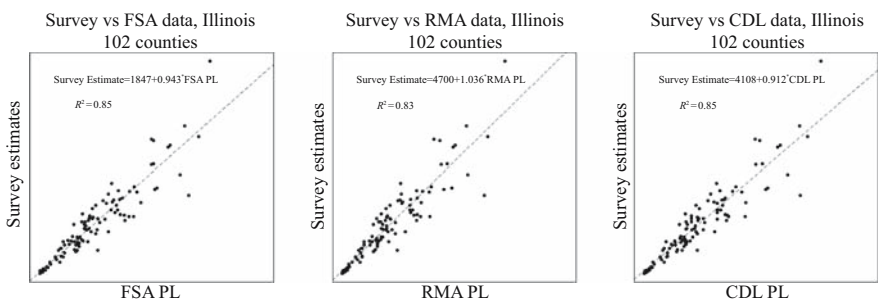


Fig. 5. Plots of survey estimates against administrative data values of planted acreage of corn available from the FSA, RMA, and CDL, respectively, overlaid with data-specific best simple linear regression lines.

prior variance, we adopt a diffuse prior for β. The prior distributions for the model variance components σ_u^2 and σ_v^2 are *Uniform*(0, 10^8) and *Uniform*(0, 10^8), respectively.

The model (3, 4, 5) borrows information from all the counties in a district and from all the districts in the state, while combining auxiliary information available at the subarea level, 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 a predictor composed of the county-level survey estimator and a composite predictor of county-level synthetic predictor and a district-level effect predictor. The derivation is provided below.

Combining (3) and (4) using Bayes' theorem, we obtain the distribution of θ_{ij} given the data and the nuisance parameters,

$$\theta_{ij}|(v_i, \beta, \sigma_u^2, \sigma_v^2, \hat{\theta}_{ij}, \hat{\sigma}_{ij}^2) \overset{ind}{\sim} N(\gamma_{ij}\hat{\theta}_{ij} + (1 - \gamma_{ij})(x_{ij}'\beta + v_i), (1 - \gamma_{ij})\sigma_u^2), \tag{6}$$

where $\gamma_{ij} = \frac{\sigma_u^2}{\hat{\sigma}_{ij}^2 + \sigma_u^2}$.

Integrating out θ_{ij} from (3) and (4), we obtain the conditional distribution of $\hat{\theta}_{ij}$,

$$\hat{\theta}_{ij}|(v_i, \beta, \sigma_u^2, \sigma_v^2, \hat{\sigma}_{ij}^2) \overset{ind}{\sim} N(x_{ij}'\beta + v_i, \hat{\sigma}_{ij}^2\sigma_u^2). \tag{7}$$

Now, combining (5) with (7) using Bayes' theorem again, we obtain the conditional distribution of v_i,

$$v_i|(\beta, \sigma_u^2, \sigma_v^2, \hat{\boldsymbol{\theta}}_i, \hat{\boldsymbol{\sigma}}_i^2) \overset{ind}{\sim} N(\gamma_i(\bar{\tilde{\theta}}_i^{\gamma} - \bar{x}_i^{\gamma'}\beta), (1 - \gamma_i)\sigma_v^2), \tag{8}$$

where $\gamma_{i.} = \sum_{j=1}^{n_i^c}\gamma_{ij}$, $\gamma_i = \frac{\sigma_u^2}{\sigma_v^2 + \sigma_u^2(\gamma_{i.})^{-1}}$, $\bar{\tilde{\theta}}_i^{\gamma} = (\gamma_{i.})^{-1}\sum_{j=1}^{n_i^c}\gamma_{ij}\hat{\theta}_{ij}$, $\bar{x}_i^{\gamma} = (\gamma_{i.})^{-1}\sum_{j=1}^{n_i^c}\gamma_{ij}x_{ij}$, $\hat{\boldsymbol{\theta}}_i$ the vector of $\hat{\theta}_{ij}$s and $\hat{\boldsymbol{\sigma}}_i^2$ is the vector of $\hat{\sigma}_{ij}^2$s.

By the conditional mean formula in (6) and (8), it follows that the posterior mean of θ_{ij}, given the data and the nuisance parameters, is

$$\mathrm{E}(\theta_{ij}|\beta, \sigma_u^2, \sigma_v^2, \hat{\theta}_{ij}, \hat{\sigma}_{ij}^2) = x_{ij}'\tilde{\beta} + \tilde{\gamma}_i(\bar{\tilde{\theta}}_i^{\gamma} - \bar{\tilde{x}}_i^{\gamma'}\tilde{\beta}) + \tilde{\gamma}_{ij}\left\{\hat{\theta}_{ij} - x_{ij}'\tilde{\beta} - \tilde{\gamma}_i(\bar{\tilde{\theta}}_i^{\gamma} - \bar{\tilde{x}}_i^{\gamma'}\tilde{\beta})\right\}, \tag{9}$$

where $\tilde{\gamma}_{ij} = \frac{\tilde{\sigma}_u^2}{\tilde{\sigma}_u^2 + \hat{\sigma}_{ij}^2}$, $\tilde{\gamma}_{i.} = \sum_{j=1}^{n_i^c}\tilde{\gamma}_{ij}$, $\tilde{\gamma}_i = \frac{\tilde{\sigma}_v^2}{\tilde{\sigma}_v^2 + \tilde{\sigma}_u^2(\tilde{\gamma}_{i.})^{-1}}$, $\bar{\tilde{\theta}}_i^{\gamma} = (\tilde{\gamma}_{i.})^{-1}\sum_{j=1}^{n_i^c}\tilde{\gamma}_{ij}\hat{\theta}_{ij}$, $\bar{\tilde{x}}_i^{\gamma} = (\tilde{\gamma}_{i.})^{-1}$ $\sum_{j=1}^{n_i^c}\tilde{\gamma}_{ij}x_{ij}$, and $\tilde{v}_i = \tilde{\gamma}_i(\bar{\tilde{\theta}}_i^{\gamma} - \bar{\tilde{x}}_i^{\gamma'}\tilde{\beta})$. The estimated variance parameters $\hat{\sigma}_u^2$ and $\hat{\sigma}_v^2$ are constructed as the posterior means for these parameters, that is $E(\sigma_u^2|\hat{\theta}_{ij}, \hat{\sigma}_{ij}^2, \beta, \sigma_v^2, \theta_{ij})$ and $E(\sigma_v^2|\hat{\theta}_{ij}, \hat{\sigma}_{ij}^2, \beta, \sigma_u^2, \theta_{ij})$, respectively.

Note that the posterior mean can be further rewritten as

$$\begin{aligned}\tilde{\theta}_{ij} &= x_{ij}'\tilde{\beta} + \tilde{\gamma}_i(\bar{\tilde{\theta}}_i^{\gamma} - \bar{\tilde{x}}_i^{\gamma'}\tilde{\beta}) + \tilde{\gamma}_{ij}\left\{\hat{\theta}_{ij} - x_{ij}'\tilde{\beta} - \tilde{\gamma}_i(\bar{\tilde{\theta}}_i^{\gamma} - \bar{\tilde{x}}_i^{\gamma'}\tilde{\beta})\right\} \\ &= \tilde{\gamma}_{ij}\hat{\theta}_{ij} + (1 - \tilde{\gamma}_{ij})\left\{x_{ij}'\tilde{\beta} + \tilde{v}_i\right\}. \end{aligned} \tag{10}$$

Using Equation (10), note the district-level contribution to the county-level not-in-sample predictions, v_i; for an area-level model, this term would be missing in Equation (10). On the other hand, Equation (10) may be rewritten as

$$\tilde{\theta}_{ij} = \tilde{\gamma}_{ij}\hat{\theta}_{ij} + (1 - \tilde{\gamma}_{ij})\, x_{ij}'\tilde{\beta} + (1 - \tilde{\gamma}_{ij})\tilde{\gamma}_i(\bar{\tilde{\theta}}_i^{\gamma} - \bar{\tilde{x}}_i^{\gamma'}\tilde{\beta}), \tag{11}$$

where $\tilde{\gamma}_{ij}$, $\tilde{\gamma}_{i.}$, $\tilde{\gamma}_i$, $\tilde{\tilde{\theta}}_i^{\gamma}$ and $\tilde{\bar{x}}_i^{\gamma'}$ are defined as for (10). The subarea-level and area-level components to the subarea-level posterior mean are clearly identified in Equation (11).

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. (2019) 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 model defined in (3), (4), and (5) to be fitted. 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 district i:

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

The counties for which data are missing in all of the data sources considered, $(\hat{\theta}_{ij}, \hat{\sigma}_{ij}^2, x_{ij})$, are excluded from the prediction set, because there is not enough evidence to conclude that planting activity took place for the specific crop, in the specific crop season. Not-in-sample predictions for these additional counties may be constructed using the methods for the third case above, after imputing covariate values x_{ij} (for example, using the average values available for other counties in the same division or state). However, not having any data to indicate county-level planting activity may lead to severe extrapolation and under-adjustments in the benchmarking step. For the cases with missing data in some of the sources, but available in others, we assume the missing at random (MAR) mechanism.

The first step in the modeling strategies is to impute the missing covariate values x_{ij}, for county j in district i, where survey estimates $(\hat{\theta}_{ij}, \hat{\sigma}_{ij}^2)$ 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,

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

over all counties k with survey and auxiliary data available. The resulting set of counties n^c with survey and auxiliary data $(\hat{\theta}_{ij}, \hat{\sigma}_{ij}^2, x_{ij})$ 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 $(\hat{\theta}_{ij}, \hat{\sigma}_{ij}^2, x_{ij})$ are available, using R JAGS (see Plummer et al. 2018), and posterior distributions are constructed using MCMC simulation. To estimate the nuisance parameters and the parameters of interest for the county-level total acreages, we use 3 chains, each of 10,000 Monte Carlo samples, 1,000 burn-in samples and thinned every nine 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; 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 β^r, $(\sigma_u^2)^r$, $(\sigma_v^2)^r$, the county-level parameters of interest θ_{ij}^r and district-level parameters of interest $\theta_i^r := \sum_{j=1}^{n_i^c} \theta_{ij}^r$, where $r = 1, \ldots, R$, and R denotes the total MCMC iterates, after burn-in and thinning, equal to 3,000 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 $(\hat{\theta}_{ij}, \hat{\sigma}_{ij}^2)$ is missing but x_{ij} is available. For this, $\left\{\theta_{ij}^r\right\}_{r=1,\ldots,R}$ are drawn from the linking model (4, 5),

$$\theta_{ij}^r | (v_i^t, \beta^r, (\sigma_u^2)^r) \overset{ind}{\sim} N(\mathrm{x}_{ij}'\beta^r + v_i^r, (\sigma_u^2)^r).$$

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. (2019). 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 district-level predictions and the district-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. (2019), 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. For this, let the state-level target be denoted by a. Then the relation

$$\sum_{i,j}^{n^{c*}} \tilde{\theta}_{ij}^B = a, \tag{12}$$

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

$$\theta_{ij,r}^B := \theta_{ij,r} \times a \times \left(\sum_{k=1}^{m} \sum_{l=1}^{n_k^{c*}} \theta_{kl,r} \right)^{-1}, \tag{13}$$

where $\theta^B_{ij,r}$ is the benchmarking-adjusted iteration, for $r = 1, \ldots, R$. Final county-level and district-level posterior summaries are constructed using the county-level iterates $\theta^B_{ij,r}$ and district-level iterates $\theta^B_{i,r} := \sum_{j=1}^{n_i^{c*}} \theta^B_{ij,r}$. 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 (13), 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 over-adjustment (under-adjustment) in the predictions.

4. Results

In this section, nationwide prediction results are presented for 2015 corn planted acreage, including a comparison of different models, model efficiency and different contributions of administrative data, serving towards the production of official statistics.

4.1. Model Comparison

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 Subsection 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 one covariate derived from FSA and RMA data (directly and imputed, when applicable); and M, the model fit to the survey data with one covariate derived from FSA, RMA and CDL data

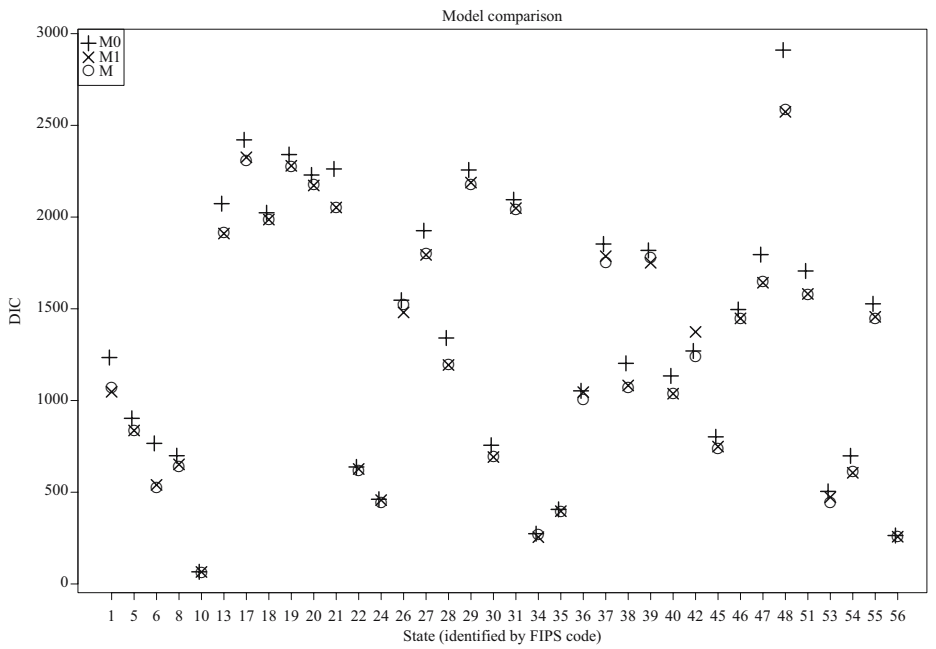


Fig. 6. Deviance information criterion (DIC) for models M0, M1, and M, by state.

Table 3. Summary of estimated factors $\tilde{\gamma}_{ij}$ (%).

Approach	Covariate ADMIN PL	1st Quantile	Median	3rd Quantile
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

(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. Also note that various data sources are combined to construct *one* covariate (models M1 and M), therefore avoiding multicollinearity issues (as is the case when multiple covariates would correspond to the data sources).

The goodness of fit for models M0, M1 and M, fit state by state, is evaluated using the Deviance Information Criterion (DIC) and results are presented in Figure 6. The x-axis in Figure 6 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 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 factors $\tilde{\gamma}_{ij}$ (%) and $\tilde{\gamma}_i$ (%) defined for (10), are constructed over all the 36 states for which the models are fit and illustrated in Tables 3 and 4. Again, model predictions constructed using M1 and M have similar features. The auxiliary data and their relationship with the survey estimates receive larger weights in the final predictions under model M compared to model M0.

4.2. Increased Number of County-Level 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 7. Model predictions are produced for 2,627 counties, of which 2,420 are in-sample counties and 207 are not-in-sample counties. Additionally, 121 model predictions were set to zero, because they corresponded to negative model predictions. 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

Table 4. Summary of estimated factors $\tilde{\gamma}_i$ (%).

Approach	Covariate ADMIN PL	1st Quantile	Median	3rd Quantile
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

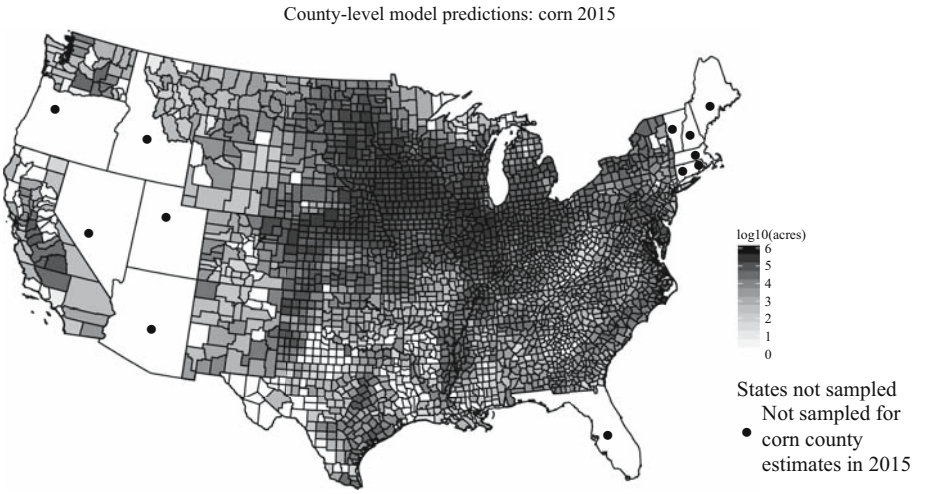


Fig. 7. Nationwide map of model-M, positive, predictions of county-level planted acreage of corn in 2015, on the log10 scale; 121 negatively-valued model predictions are set to zero.

prediction is approximately 60% the median of the in-sample model predictions) and large CVs. In contrast, recall that survey estimates are available for 2,426 counties, as illustrated in Figure 1, and under model M1, 2,486 model predictions are produced.

4.3. 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 8, 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),

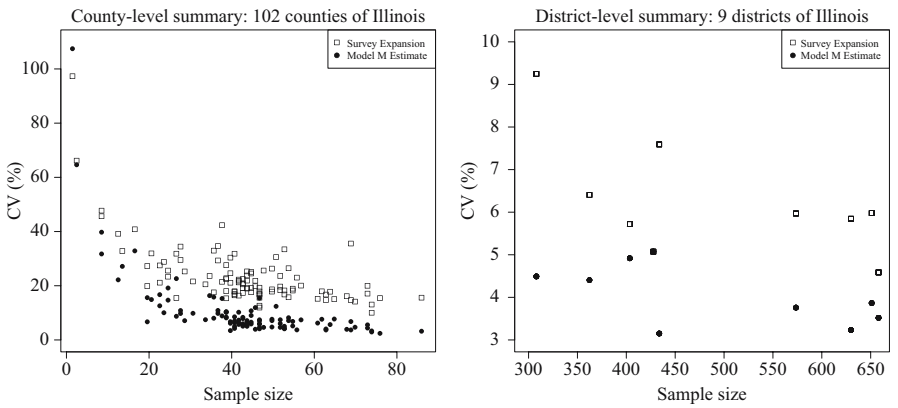


Fig. 8. Plots of CVs of survey estimates and model-M predictions of planted acreage of corn in 2015 against sample size, in Illinois, at the county level and district level.

Table 5. Summaries of standard errors of county-level survey estimates and model predictions (acres) Counties with available survey estimates.

Approach	Covariate ADMIN PL	1st Quantile	Median	3rd Quantile
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

comparing the district-level survey SEs (CVs) to the model SEs (CVs) for models M1 and M. Comparing a model's performance versus a 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 under the two different models (M and M1, respectively).

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 9, in the first row for 2,420 counties with positive survey estimates and model predictions, and in the second row for the corresponding 272 districts (which may include additional model predictions); counties or districts with relative efficiency values greater than 3 are removed to facilitate visualization. The relative SE (CV) 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.

5. Discussion

In this article, we illustrated the contributions of administrative data to produce agricultural official statistics. 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 other years (2014, 2015, and 2016), and for commodities (corn, soybean, and sorghum). 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. The number of positive model

Table 6. Summaries of standard errors of district-level survey estimates and model predictions (acres).

Approach	Covariate ADMIN PL	1st Quantile	Median	3rd Quantile
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

Table 7. Summaries of CVs (%) of county-level survey estimates and model predictions Counties with available survey estimates.

Approach	Covariate ADMIN PL	1st Quantile	Median	3rd Quantile
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

predictions is larger than the number of available survey estimates. As another external validation exercise, we compared the model predictions and the corresponding official values, for the counties and districts where both were available, 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).

Our first contribution is a novel use of administrative data to determine the set of subareas with crop-specific planting activity. 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 Native American children aged 5–17 in poverty makes sense only for domains where at least one Native American child aged 5–17 lives).

In order to construct the prediction space, we assume that the data sources considered exhaust the information available on planting activities, for a specific crop, in a specific year. However, exploration of additional sources of data is of interest. When available, such additional information (state-specific, commodity-specific and time-specific) may be used to redefine the set of subareas for which model predictions are to be constructed and to redefine the set of covariates. Also, we acknowledge, but have to ignore the possible errors in administrative planting acreage values. One extension to deal with the possible downward bias in FSA, RMA, and CDL would be to adjust the model to

$$\hat{\theta}_{ij}|\theta_{ij}, \kappa_{ij} \overset{ind}{\sim} N(\kappa_{ij}\theta_{ij}, \hat{\sigma}^2_{ij}),$$
$$\kappa_{ij} \overset{ind}{\sim} \text{Uniform}(1, a_0), \quad \theta_{ij}|v_i, \beta, \sigma^2_u \overset{ind}{\sim} N(\mathrm{x}'_{ij}\beta + v_i, \sigma^2_u),$$
$$v_i|\sigma^2_v \overset{ind}{\sim} N(0, \sigma^2_v),$$

with the same priors adopted for the parameters $(\beta, \sigma^2_u, \sigma^2_v)$, a multiplicative offset κ_{ij} and a prespecified constant a_0, say between 1 and 1.1.

Table 8. Summaries of CV(%) of district-level survey estimates and model predictions.

Approach	Covariate ADMIN PL	1st Quantile	Median	3rd Quantile
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

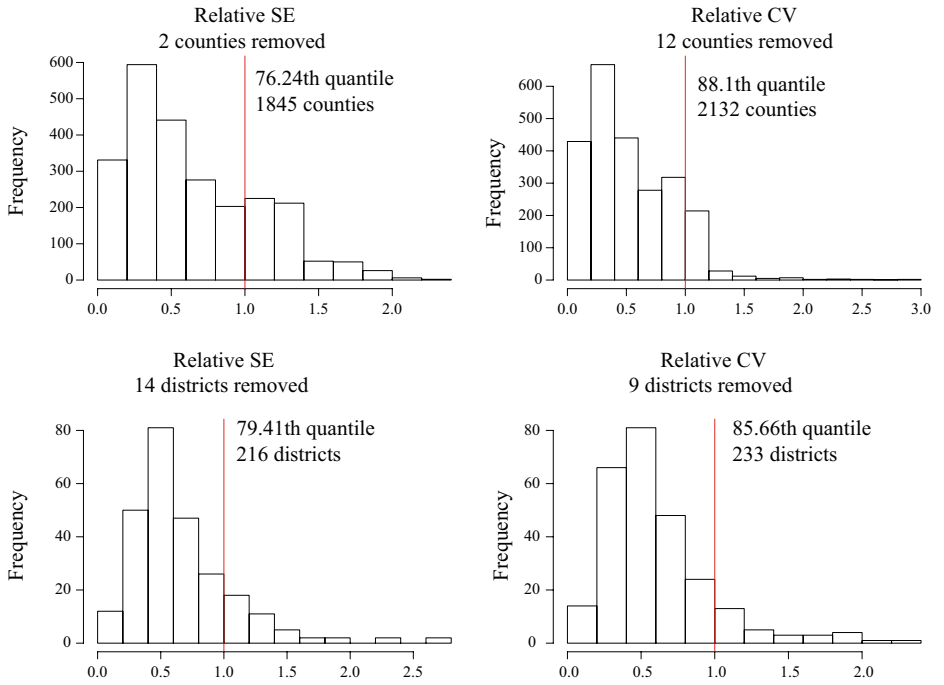


Fig. 9. Histograms of relative standard errors and relative CVs at the county level and district level, model versus survey. Relative efficiency values greater than 3 are removed to facilitate visualization.

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 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 are of interest for future research.

As a consequence of the model specification, in particular the normality assumption in the linking 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 (or preferably generalized gamma distribution, lognormal being a special case), we recognize the simplicity of the current specification, especially with respect to prediction and benchmarking at multiple levels of interest. Under a non-normal distribution, the model predictions would need to be back-transformed. This additional operation would have to be performed at the lowest level of aggregation (for our application, the county), and followed by benchmarking adjustments and aggregations to higher levels of interest.

The models were applied separately, for each state, in order to follow with the current NASS process of constructing official statistics; results are communicated to each state individually, and final dissemination follows. One may extend the model to using a three-fold model by including an additional random effect corresponding to the states, and by using the

nation-wide data. On the other hand, careful validation may be conducted at the state-level and specific auxiliary data, in addition to the ones considered here, may be incorporated.

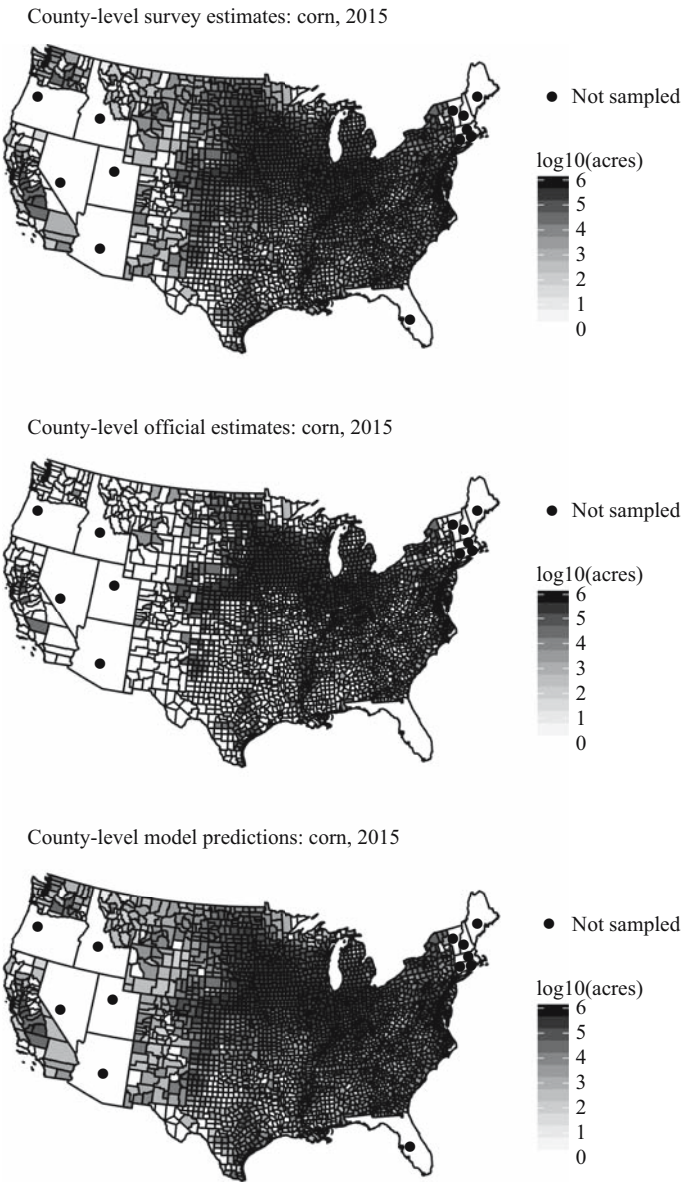
Increasing the number of counties with planted acreage predictions is another important contribution. For corn in 2015, the largest number of not-in-sample predictions happens to be in Texas: 42 out of 184 counties, accounting for approximately 0.7% of the total planted acreage. See the Appendix (Section 6) 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 over-adjusting 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.

Finally, a 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. Addressing challenges associated with the publication process is an ongoing area of interest. 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 the National Academies of Sciences, Engineering, and Medicine (2017, 117) for more details. For this application study, we investigate a hypothetical CV-based assessment, consistent with the publication standards at other government agencies (Marker 2015 reported CV-based assessments used by various government agencies). Using a 30% threshold for the county-level CVs across the nation leads to 1,694 candidate county-level planted acreage predictions for publication of corn in 2015; see Figure 10 in the Appendix (Section 6). In contrast, in 2015, NASS published estimates of corn for 1,433 counties, which are available in NASS QuickStats (USDA NASS 2016b). Moreover, in Equation (10), 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 a 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. The current NASS publication standard is being revised; see Cruze et al. (2018) for recent research on this topic.

6. Appendix: Increased Number of Reliable Estimates for Other Commodities

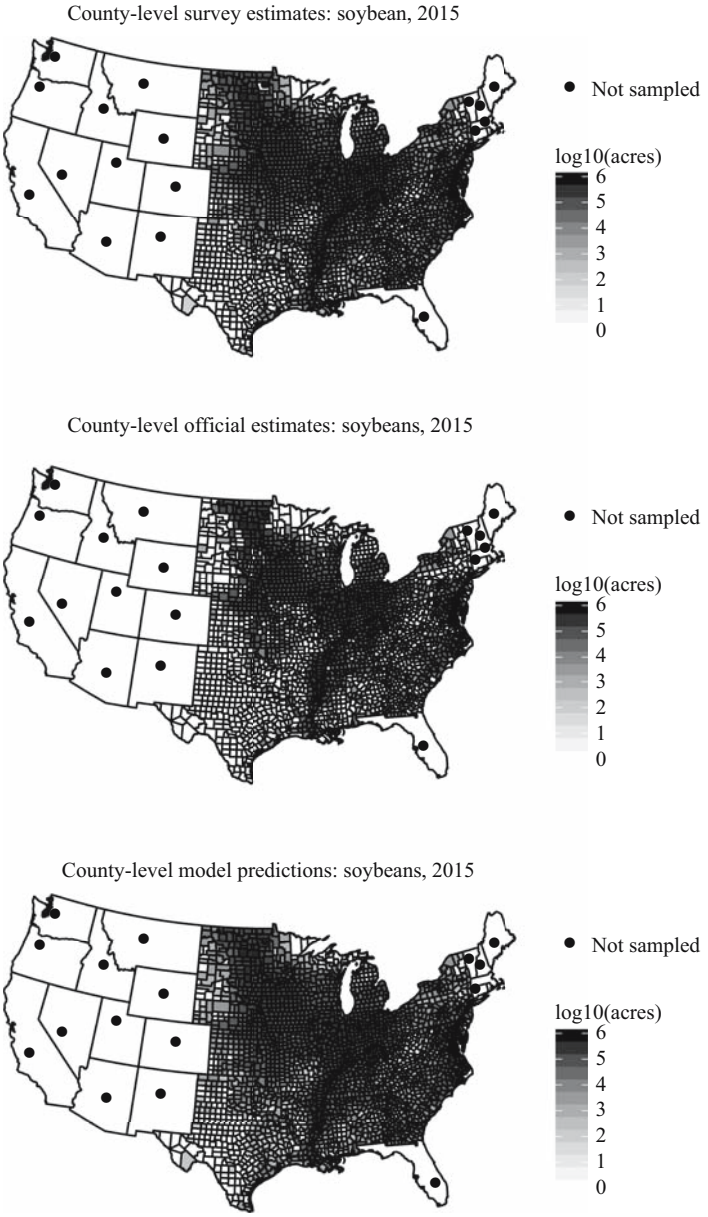
For corn and soybean in 2015, the largest numbers of not-in-sample predictions are, respectively, 42 and 70 out of 184 and 122 counties, accounting for approximately, respectively, 0.7% and 11.83% of the total planted acreage in Texas. The largest numbers of not-in-sample predictions for sorghum and winter wheat in 2015 are, respectively, 28 and 38 out of 73 and 154 counties, accounting for approximately, respectively, 5.23% and 12.47% of the total planted acreage in Mississippi and Georgia, respectively.

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.



- 1,433 official values
- 2,426 survey estimates; 1,125 have CVs ≤ 30%
- 2,627 model predictions; 1,694 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

Fig. 10. Nationwide maps of survey estimates, official values, and model-M predictions of county-level planted acreage of corn in 2015, on the log10 scale.



- 1,306 official values
- 2,012 survey estimates; 1,046 have CVs ≤ 30%
- 2,224 model predictions; 1,472 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

Fig. 11. Nationwide maps of survey estimates, official values, and model-M predictions of county-level planted acreage of soybean in 2015, on the log10 scale.

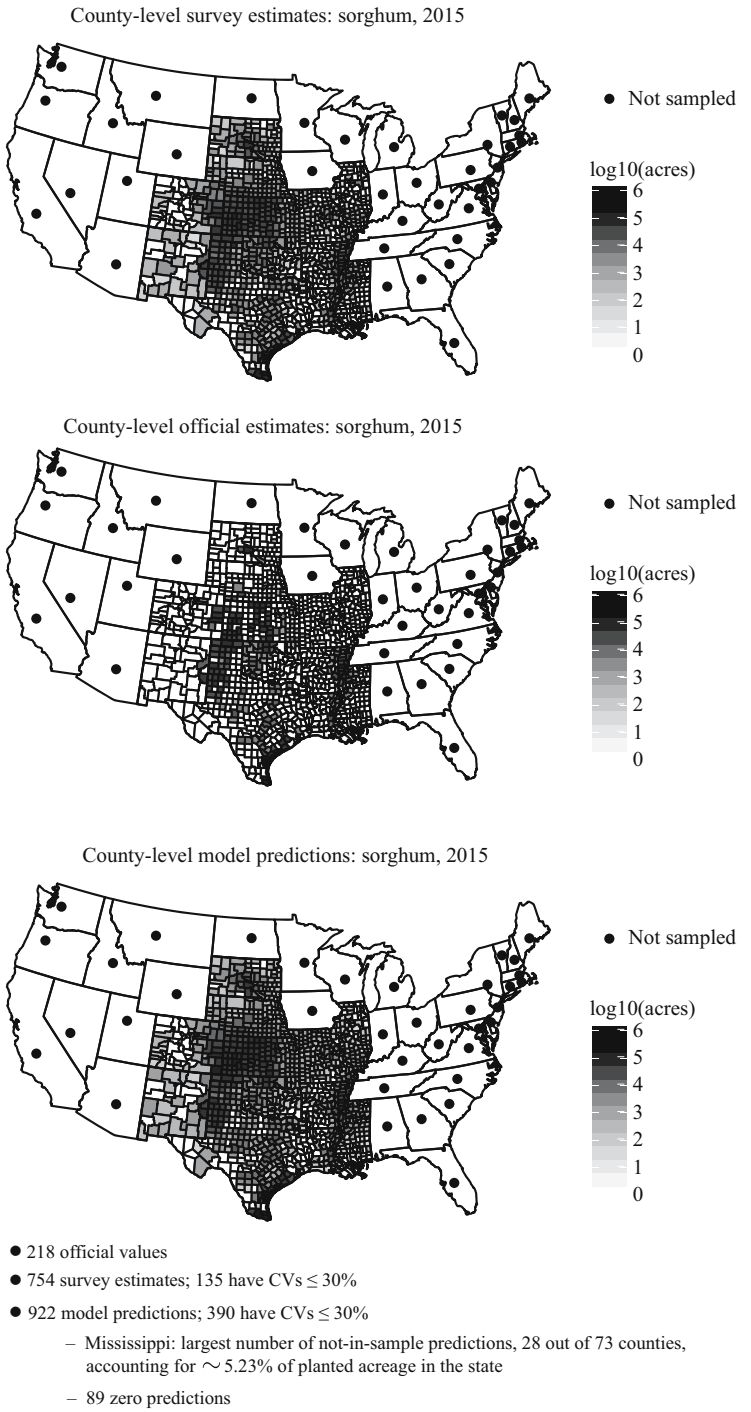


Fig. 12. Nationwide maps of survey estimates, official values, and model-M predictions of county-level planted acreage of sorghum in 2015, on the log10 scale.

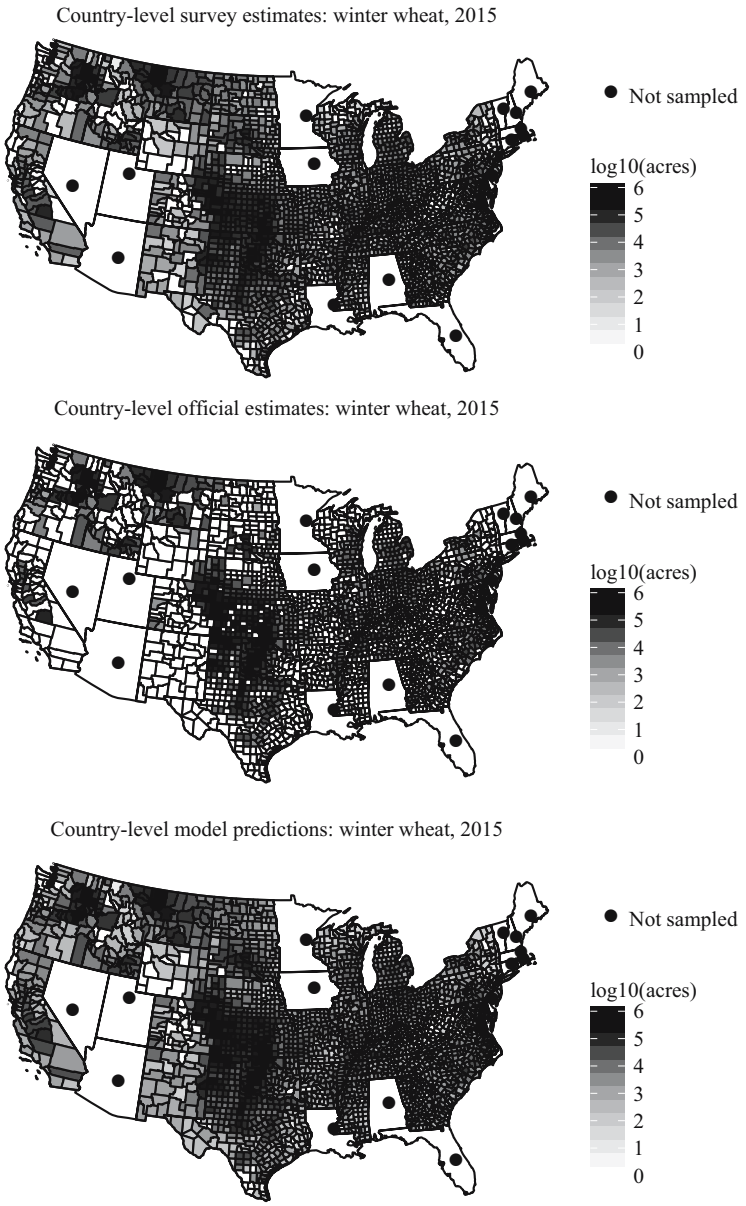


Fig. 13. Nationwide maps of survey estimates, official values, and model-M predictions of county-level planted acreage of winter wheat in 2015, on the log10 scale.

7. References

Boryan, C., Z. Yang, R. Mueller, and M. Craig. 2011. "Monitoring US agriculture: the US Department of Agriculture, National Agricultural Statistics Service, Cropland Data Layer Program." *Geocarto International* 26(5): 341–358. DOI: https://doi.org/10.1080/10106049.2011.562309.

Cruze, N.B., A.L. Erciulescu, B. Nandram, W.J. Barboza, and L.J. Young. 2019. "Producing Official County-Level Agricultural Estimates in the United States: Needs and Challenges." *Statistical Science* 34(2): 301–316. DOI: https://doi.org/10.1214/18-STS687.

Cruze, N.B., A.L. Erciulescu, H. Benecha, V. Bejleri, B. Nandram, and L.J. Young. 2018. "Toward an Updated Publication Standard for Official County-Level Crop Estimates." *Joint Statistical Meetings Proceedings. Government Statistics Section*. Alexandria, VA: American Statistical Association. 1576–1585. Available at: https://www.nass.usda.gov/Education_and_Outreach/Reports,_Presentations_and_Conferences/reports/conferences/JSM-2018/Toward_an_updated_publication_standard_for_official_county-level_crop_estimates.pdf (accessed September 2019).

Erciulescu, A.L., N.B. Cruze, and B. Nandram. 2018. "Benchmarking a Triplet of Official Statistics." *Environmental and Ecological Statistics* 25: 523–547. DOI: https://doi.org/10/1007/s10651-018-0416-4.

Erciulescu, A.L., N.B. Cruze, and B. Nandram. 2019. "Model-Based County-Level Crop Estimates Incorporating Auxiliary Sources of Information." *Journal of the Royal Statistical Society, Series A* 182: 283–303. DOI: https://doi.org/10/1111/rssa.12390.

Fay, R.E. and R.A. Herriot. 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. DOI: https://doi.org/10.1080/01621459.1979.10482505.

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

Gelman, A. and D.B. Rubin. 1992. "Inference from iterative simulation using multiple sequences." *Statistical Science* 7: 457–511. DOI: https://doi.org/10.1214/ss/1177011136.

Geweke, J. 1992. "Evaluating the accuracy of sampling-based approaches to calculating posterior moments." In *Bayesian Statistics 4*, edited by J.M. Bernado, J.O. Berger, A.P. Dawid, and A.F.M. Smith. Oxford, UK: Clarendon Press.

Kennedy, C., A. Mercer, S. Keeter, N. Hatley, K. McGeeney, and A. Gimenez. 2016. "Evaluating Online Nonprobability Surveys," Pew Research Center. Available at: https://www.pewresearch.org/methods/2016/05/02/evaluating-online-nonprobability-surveys/ (accessed September 2019).

Kim, J.K., Z. Wang, Z. Zhu, and N.B. Cruze. 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. DOI: https://doi.org/10.1007/s13253-018-0320-2.

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 (accessed September 2019).

National Academies of Sciences, Engineering, and Medicine. 2017. "Improving Crop Estimates by Integrating Multiple Data Sources," Washington, DC: The National Academies Press. DOI: 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. Available at: https://www.nass.usda.gov/Education_and_Outreach/Reports,_Presentations_and_Conferences/Survey_Reports/Estimating\%20the\%20Coverage\%20of\%20Farm\%20Service\%20Agency\%20Crop\%20Acreage\%20Totals.pdf (accessed September 2019).

Plummer, M., A. Stukalov, and M. Denwood. 2018. "Bayesian Graphical Models using MCMC," Version 4–8. Available at: https://cran.r-project.org/web/packages/rjags/rjags.pdf (accessed September 2019).

Rao, J.N.K. and I. Molina. 2015. *Small Area Estimation*. 2nd ed. Hoboken: Wiley.

Torabi, M. and J.N.K. Rao. 2014. "On small area estimation under a sub-area level model." *Journal of Multivariate Analysis* 127: 36–55. DOI: https://doi.org/10.1016/j.jmva.2014.02.001.

USDA FSA. 2019. "United States Department of Agriculture Farm Service Agency: ARC/PLC Program." Available at: https://www.fsa.usda.gov/programs-and-services/arcplc_program/index (accessed September 2019).

USDA NASS. 2016a. "CropScape and Cropland Data Layer." Available at: https://www.nass.usda.gov/Research_and_Science/Cropland/SARS1a.php (accessed September 2019).

USDA NASS. 2016b. "QuickStats." Available at: https://quickstats.nass.usda.gov/ (accessed September 2019).

USDA NASS APS. 2018. "Crops/Stocks Agricultural Survey." Available at: https://www.nass.usda.gov/Surveys/Guide_to_NASS_Surveys/Crops_Stocks/index.php (accessed September 2019).

USDA NASS CAPS. 2018. "County Agricultural Production." Available at: https://www.nass.usda.gov/Surveys/Guide_to_NASS_Surveys/County_Agricultural_Production/index.php (accessed September 2019).

USDA NASS CDL. 2018. "CropScape and Cropland Data Layers – FAQs." Available at: https://www.nass.usda.gov/Research_and_Science/Cropland/sarsfaqs2.php (accessed September 2019).

USDA RMA. 2019. "United States Department of Agriculture Risk Management Agency: FCIC." Available at: https://www.rma.usda.gov/Federal-Crop-Insurance-Corporation (accessed September 2019).

Received October 2018
Revised May 2019
Accepted September 2019