Developments In Model-Based Estimation Of County-Level Agricultural Estimates

Nathan B. Cruze[1] Andreea L. Erciulescu[2,1] Balgobin Nandram[3,1]
Wendy J. Barboza[1] Linda J. Young[1]

Abstract
In the United States, county-level estimates of crop yield, production, and acreage published by the United States Department of Agriculture's National Agricultural Statistics Service (USDA NASS) play an important role in determining the value of payments allotted to farmers and ranchers enrolled in several Federal programs. Given the importance of these official small area crop estimates, NASS continually strives to improve the county estimates program in terms of accuracy, reliability, and coverage. NASS is pursuing research in model-based estimation of county crop acreage and yield estimates as a viable strategy for combining several existing and potential sources of survey, administrative, and auxiliary data. In 2015, NASS engaged a panel of experts convened under the auspices of the National Academies of Sciences, Engineering, and Medicine Committee on National Statistics (CNSTAT) for guidance and collaboration on models that may synthesize multiple sources of information into a single estimate, provide reasonable measures of uncertainty, and potentially increase the number of publishable county estimates. This paper describes developments in NASS research in model-based county crop estimates since the inception of the County Agricultural Production Survey (CAPS) in 2011. Some of the ongoing needs and research challenges are noted.

Key Words: agricultural surveys, auxiliary data, benchmarking, official statistics, small area estimation

1. Introduction

In the United States, county-level agricultural estimates of crop acreage, production, and yield published by USDA's National Agricultural Statistics Service (NASS) play a pivotal role in the administration of farm subsidy, crop insurance, and agricultural support programs. The Congressional Budget Office projects that the cost of administering the programs in the Agricultural Act of 2014 (the 2014 Farm Bill) will total $489 billion from 2014 to 2018 (Congressional Budget Office 2014). Outlays for crop insurance, conservation, and commodity programs will account for nearly one-fifth of those expenditures. Two USDA agencies play critical roles in the administration of those programs: the Farm Service Agency (FSA) and the Risk Management Agency (RMA). Both FSA and RMA rely on NASS county estimates as important thresholds or benchmarks as they disburse funds through their respective agricultural programs. In short, NASS crops county estimates may *determine* the magnitude of disbursements individual farmers receive under several farm programs.

Given the importance of these official small area crop estimates, NASS continually strives to improve the county estimates program in terms of accuracy, reliability, and coverage. NASS uses the term 'indication' to refer to any input information source and reserves the term 'estimate' for an official statistic. Traditionally, NASS official crop esti-

[1] United States Department of Agriculture National Agricultural Statistics Service, 1400 Independence Ave, SW, Washington, DC 20250

[2] National Institute of Statistical Sciences, Washington Square, 1050 Connecticut Avenue NW, Washington, DC 20036

[3] Worcester Polytechnic Institute, Department of Mathematical Sciences, Stratton Hall, 100 Institute Road, Worcester, MA 01609

mates at any geographic level (national, state, agricultural statistics district, and county) are an expert-weighted assessment of several *indications* as determined by the Agricultural Statistics Board (ASB). These indications could be obtained from NASS surveys, additional administrative data, or other forms of auxiliary information. Recognizing the need to strengthen the probability sample underlying all sub-state survey indications, NASS fully implemented the County Agricultural Production Survey (CAPS) in 2011 as a complementary survey to augment the samples obtained under existing quarterly surveys. Despite taking this step, small sample sizes may still be realized for some commodities in some counties due to nonresponse, changes in respondents' year-to-year planting decisions, and the sparsity of certain crops within given administrative boundaries. Moreover, the increase in availability of auxiliary, non-survey crop data including administrative sources, remote sensing, and weather data motivate the need for model-based approaches to estimate small area agricultural estimates. Consequently, NASS has renewed its interest in small area estimation and model-based approaches to combining survey data with other available sources of information. In particular, NASS has convened an expert panel under the auspices of the National Academies of Sciences, Engineering, and Medicine Committee on National Statistics (CNSTAT) to review its county estimates program and to advise on the possible role model-based estimates could serve in the production of official statistics.

This paper discusses some of the challenges and needs of the NASS crops county estimates program. The NASS survey cycle and input data available for sub-state level estimates are discussed in Section 2. In Section 3, a preliminary case study for planted acreage is used to highlight some of the ongoing questions NASS faces as it considers the increased use of model-based estimation in the production of its official statistics. Discussion and conclusions are offered in Section 4.

2. NASS Survey Cycle, Data, and County Estimates

2.1 Survey Cycle

NASS has the challenging task of estimating planted and harvested area, production, and yield for the diverse array of crops grown across the entire United States. For many of its surveys, NASS implements a multivariate probability proportional to size (MPPS) survey design, which offers additional flexibility to target key crops grown within each state (Bailey and Kott 1997). A partial NASS survey cycle and publication timeline is depicted in Figure 1; the width of each interval represents the approximate data collection window for each survey. NASS conducts quarterly Acreage, Production, and Stocks (APS) surveys in an ongoing effort to capture activities throughout the life cycle of the crop, including intended planting decisions (March), indications of planted acreage (June), and indications of harvest and output activities for small grains crops (September) and major row crops (December). An area frame component of the June survey provides an undercoverage adjustment for the list-based samples obtained during the September and December APS surveys. Coverage-adjusted national and state survey indications are available to the ASB for their deliberations. The official ASB consensus estimates for state and national activity are released in the Small Grains Summary in late September or in the Annual Summary (for row crops) in January of the following calendar year.

As shown in the timeline, the data collection window for the CAPS surveys, from questionnaire mail out to final summary, extends beyond the release of the national and state official statistics; *official estimates for state acreages, production, and yield are determined prior to the publication of any county estimates.* Within the CAPS data collection window, NASS applies adaptive techniques to increase coverage with respect to targeted commodi-

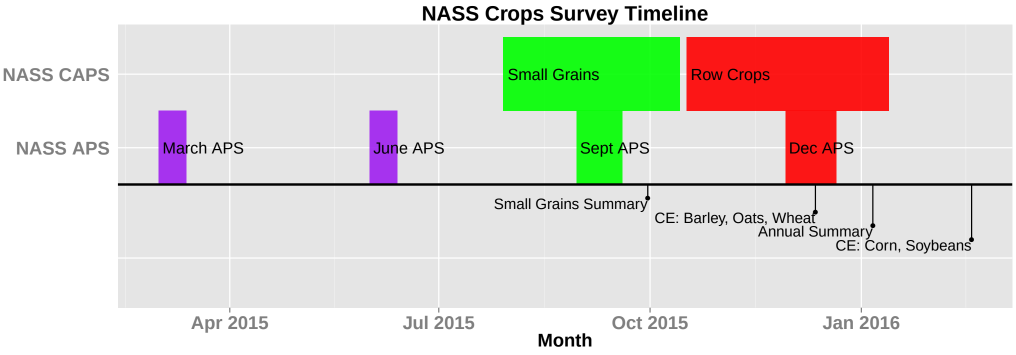


Figure 1: NASS crop survey and publication timeline

ties. At the conclusion of CAPS data collection, the MPPS samples of both the APS survey and CAPS are pooled and reweighted. The county-level survey indications are, in effect, computed from a single *list-based* sample. Acquisition of additional observations during CAPS data collection can potentially more than double the total number of responses obtained within each state relative to the APS survey sample size alone, however, this does not guarantee that the number of reports *for each commodity* will double as the sampled respondents may not grow all types of crops that NASS may seek to estimate. The official county estimates for small grains (e.g., barley, oats, and wheat) are published in December. The first row crops county estimates for corn and soybeans are published in February of the following calendar year. Row crops county estimates for additional commodities are subsequently released at intervals into the month of May.

NASS conducts the row crops CAPS surveys in 43 states excluding the 5 New England states shown in red in Figure 2. The small grains CAPS is also conducted in the group of 37 states shown in blue. The list of commodity crops targeted may differ from state to state and from year to year, subject to providing required coverage for federally mandated program crops, and satisfying the needs of other stakeholders, e.g., specific state program commodities.

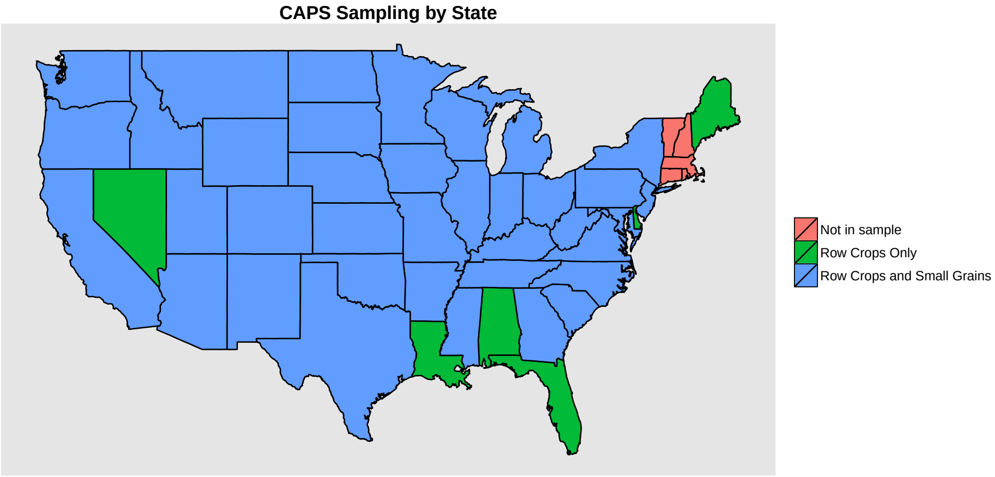
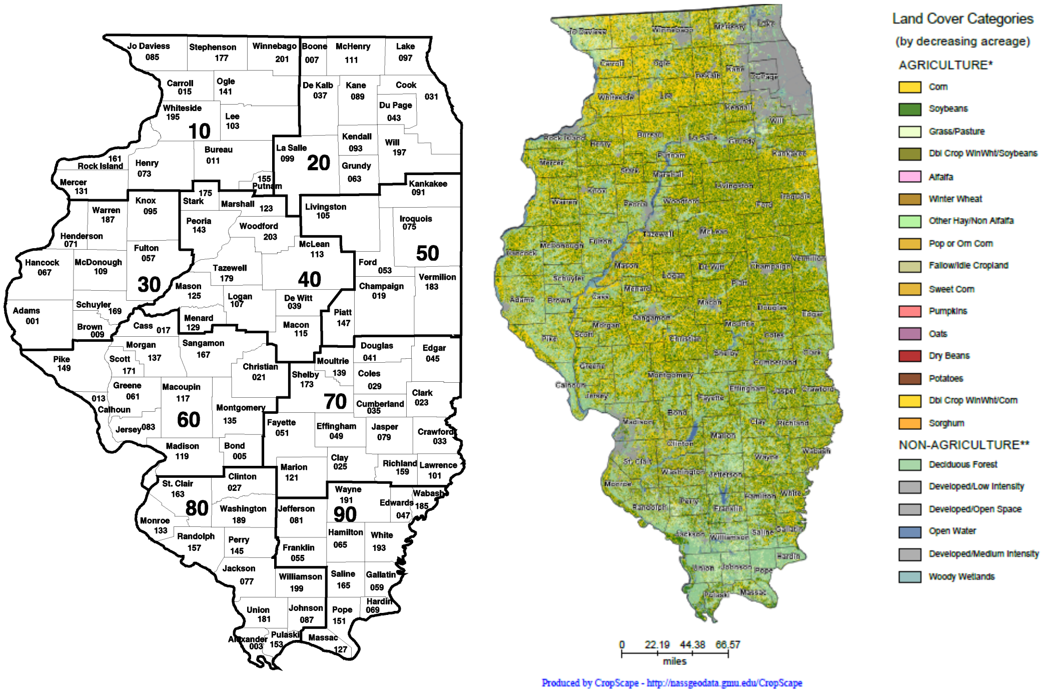


Figure 2: Row crops and small grains CAPS states

2.2 Survey Data, Administrative Data, and Other Auxiliary Data

Although the county is generally the smallest area estimate and most desirable level of estimate for the administration of policy, NASS also produces estimates at an intermediate domain between the county and state called the agricultural statistics district (ASD). The ASD is a predefined group of neighboring counties within a state. Small states may have only a single ASD, whereas Texas has 15 ASDs. More typically, states will have 9 or 10 ASDs. NASS utilizes its own survey data, administrative data provided by FSA and RMA, and remote sensing data available at both the county level and the ASD level when setting official sub-state crop estimates.

Figure 3 depicts the division of the state of Illinois into its 9 ASDs comprised of 102 counties. The right panel shows the corresponding Cropland Data Layer (CDL), a remote sensing land-cover classification product produced by NASS (Boryan 2011). This figure illustrates the diversity of crop cover that may be grown in the state throughout the year, and it highlights the challenges of detecting and sampling a variety of commodities that may not be widely grown throughout all parts of the state.



Produced by CropScape - http://nassgeodata.gmu.edu/CropScape

Figure 3: Maps of Illinois ASD and county boundaries and corresponding CDL land cover

In addition to remote sensing data, NASS obtains FSA administrative data at the county level. Farmers who wish to participate in FSA programs certify the acreages and crop types that they grow with the FSA. Since FSA programs are voluntary, the FSA acreage data may have some degree of undercoverage; these data represent a minimum amount of planting activity that is known to have taken place within the county lines, and it is essentially thought of as a lower bound for county estimates of planted area. The extent of undercoverage in the FSA data may differ by state and even by commodities within state.

As an underwriter of crop insurance policies, the RMA receives administrative data on *failed acreage* (acreage that was planted but not harvested for any reason) from various independent crop insurance agents. Again, the RMA administrative data likely represent a minimum amount of crop abandonment as not all farmers will participate in a crop insurance program, and even those who do may not file a claim in the given crop year.

Little administrative data exist to help measure total agricultural output or yield at the county level. Remote sensing approaches to measuring yield are increasingly of interest to NASS and to the agricultural sector at large (Johnson 2014; Johnson 2016). In time, new technologies such as precision agricultural instruments on board tractors and combines may help inform estimates of production and yield. Currently, data ownership and privacy issues for precision agricultural data are still being debated and developed at the Congressional and Departmental levels (Hurst 2015).

2.3 The Official Statistics

2.3.1 Sub-state Crop Estimates

At both the ASD and county level, NASS produces estimates of four parameters of interest: total planted area, total harvested area, total production, and yield. Since the state estimates are determined prior to setting county estimates, NASS employees a ‘top-down’ strategy, first setting ASD-level estimates for acreage and production subject to Equation 1 and then setting corresponding county estimates subject to Equation 2.

$$Total_{State} = \sum_{ASD \in State} Total_{ASD} \tag{1}$$

$$Total_{ASD} = \sum_{counties \in ASD} Total_{county} \tag{2}$$

Within each administrative boundary, estimates of yield are obtained by dividing total production estimates by corresponding total harvested acreage estimates.

Letting **PL, HV, PD** and **YD** denote official estimates of planted area, harvested area, production, and yield, respectively, Table 1 summarizes the process by which a commodity expert would first set ASD-level estimates for all four parameters of interest followed by county-level estimates of the same parameters. Beginning with planted area, a commodity statistician reviews the ASD-level survey indication, administrative acreage data obtained from FSA, and, where available, a Battese-Fuller model-based indication of planted area. Note that the latter is a *unit-level nested error regression model* (Fuller and Battese 1973) for estimating small areas which adjusts satellite pixel count data using NASS June Area Survey data as a source of ground truth (Walker and Sigman 1982). Use of the model as a separate indication rather than the official statistic has been typical of the NASS paradigm to date. Although participation in FSA programs is popular, it is not compulsory, and some degree of undercoverage may exist in the FSA data. Commodity statisticians set **PL** estimates so as to cover this minimum level of planting activity whenever possible. Indeed, reconciling known relationships among the input data sources is an important part of what the expert must accomplish when setting the estimate.

Once the expert has established ASD-level estimates for **PL** satisfying Equation 1, he may proceed to setting ASD-level estimates for harvested area. Two current year survey indications may be used to help inform harvested area estimates: a survey-based total as well a survey-based ratio of harvested to planted area multiplied by **PL** determined in the previous step. Additionally, an indication derived from RMA administrative acreage data, and the past year’s harvested area estimate are assessed. Current crop year estimates of harvested acreage (**HV**) must be derived satisfying Equation 1 and subject to the additional constraint that **HV**$\leq$**PL** within every area of interest.

In a similar manner, all ASD-level estimates of production (**PD**) are set subject to Equation 1, based on an assessment of survey indications, past estimates, and remote sensing indications. Note that the remote sensing indications are only available for corn and soybeans in a limited number of states, and they do not incorporate any survey information at this time; see Johnson (2014) for details. Little exists in the way of administrative data for production. The assessment of distinct indications takes place on the production scale; the corresponding yield estimates are derived as the ratio of **PD** and **HV** estimates.

Table 1: Summary of indications reviewed for ASD and county crop estimates. **Denotes indications which may not be available for all commodities or within every state

	Planted Area	*Harvested Area*	*Production/Yield*
Indications	Survey Planted Total FSA Certified Acreage Remote Sensing (Battese-Fuller)**	Survey Harvested Total **PL**×Survey Harvested/Planted Ratio **PL**–RMA Failed Acreage Previous Year **HV**	Survey Production Total **HV**×Survey Yield Ratio **HV**×Remote Sensing Yield** Previous Year **PD**
ASD Estimates	$\mathbf{PL}_{\mathbf{ASD}}$	$\mathbf{HV}_{\mathbf{ASD}}$	$\mathbf{PD}_{\mathbf{ASD}}$ and $\mathbf{YD}_{\mathbf{ASD}}$
County Estimates	$\mathbf{PL}_{\mathbf{county}}$	$\mathbf{HV}_{\mathbf{county}}$	$\mathbf{PD}_{\mathbf{county}}$ and $\mathbf{YD}_{\mathbf{county}}$

Once all ASD-level estimates for a particular commodity have been set, county estimates are set in an analogous manner, satisfying Equation 2 for each total. Again, county yields are derived as a consequence of setting the **PD** and **HV** totals.

The process of manual assessment of separate indications is time consuming, and it must be repeated for each state and commodity separately. Table 1 shows that the estimates of harvested area are functions of the planted area estimates in the previous step, and the production estimates are a function of the harvested estimates. Any errors may be propagated through this sequence and down to lower spatial scales; quantification of the uncertainty associated with the official estimates produced in this manner is difficult. *Presently, NASS does not publish any measures of uncertainty for its sub-state crop estimates.* Presumably, weather, drought, and soil information may be informative for estimates of production and yield. Aside from the remote sensing yield indication, there is no indication that translates any changes due to these factors directly onto the production scale for use by the commodity expert. An appropriate model-based strategy could enhance reproducibility, quantify associated uncertainties, and enforce benchmarking constraints while incorporating a variety of data types.

2.3.2 The Current Publication Standards

Although NASS does not publish its survey outcomes directly, the features of the survey indications determine whether an estimate at the county or ASD level can be published. NASS uses a compound rule for its publication standard, verifying either:

- a minimum sample size, or
- a minimum area coverage threshold.

Item-level nonresponse is permitted for production and yield, meaning that a respondent taking the survey could provide acreage information but decline to provide his total production or yield. The number of reports that determine production and yield may be smaller than the number of reports used to estimate acreages. NASS will first check that at least 30 valid positive reports of production or yield (respondents may report either quantity on the questionnaire) have been realized within the county, in which case the county estimate may be published. If 30 positive production or yield reports cannot be realized in the county for the commodity of interest, a second check for coverage is made to determine whether county estimates can still be published. A harvested area expansion *based on the (possibly smaller) number of realized yield reports* is compared to the harvested area *estimate* (the official statistic determined by the ASB); county estimates can still be published provided a minimum of 25% coverage is obtained based on at least 3 positive yield reports. NASS will either publish estimates of all parameters with respect to each commodity in the county, i.e., planted and harvested area, production, and yield, or it will suppress all estimates of that commodity. Counties that must be suppressed will be grouped into larger aggregates of counties within the state, and those aggregates must also pass the publication standard. These rules are applied for each commodity independently; it may be possible to publish estimates for one commodity in a given county, but necessary to suppress estimates for a different crop within the same county.

Figure 4 shows publication and suppression of NASS county estimates of corn for the 2014 crop year. Colored counties indicate counties where some amount of corn planting has been indicated within the county. Ideally, all of these counties would be supported by official estimates. Green counties indicate those counties that NASS was able to publish according to its publication standard. Counties shown in red are those counties that failed either part of the compound publication standard. Since the state total is known, it would be possible in some cases to back-calculate unpublished counties given the published counties. Complementary suppression, shown in yellow, is sometimes a necessary step. The result is that county estimates that may otherwise be suitable for publication are suppressed for reasons entirely outside the boundary of that county.

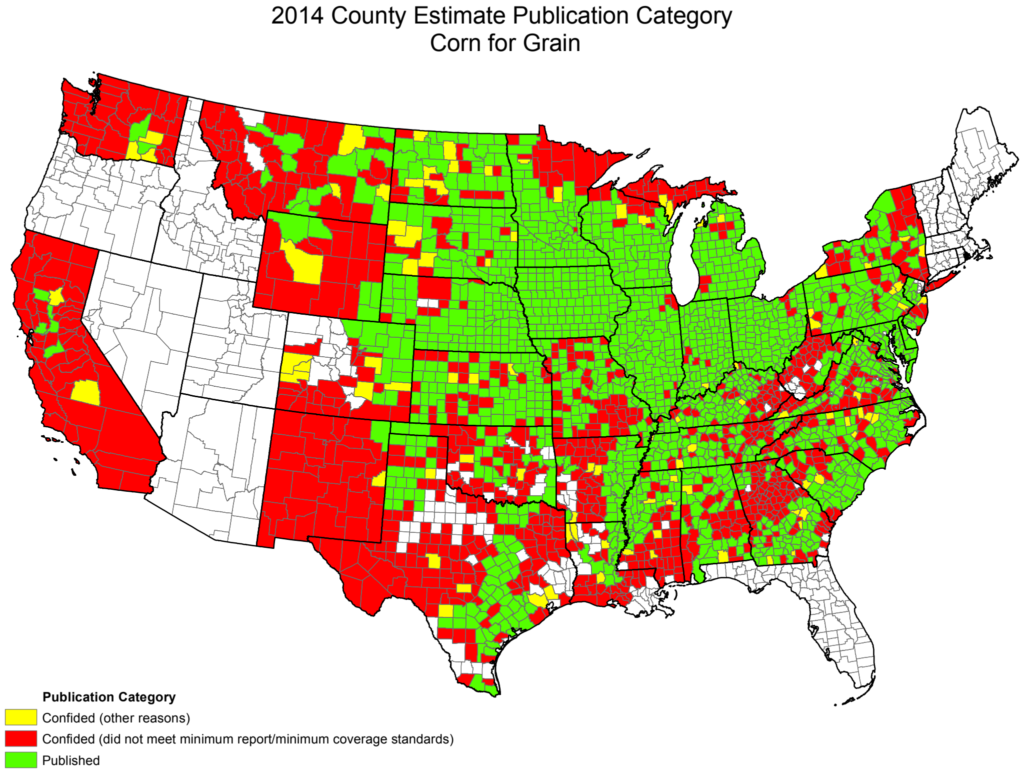

2014 County Estimate Publication Category
Corn for Grain
Publication Category
🟨 Confided (other reasons)
🟥 Confided (did not meet minimum report/minimum coverage standards)
🟩 Published

Figure 4: Map of published and confided county estimates of corn in crop year 2014

Even when NASS suppresses county estimates, the FSA and RMA must continue to administer programs in those counties. Although NASS produces aggregates of several counties or ASD-level estimates, those estimates may lack the specificity required to administer good local farm policy. Model-based approaches may help stabilize small area crop estimates. NASS is currently considering what additional roles model-based estimates could play in the production of *official statistics* at the county and ASD levels in order to increase the number of published county estimates and better support other USDA agencies in their respective missions. An important part of this deliberation is to determine how the availability of a model-based estimate could inform or function within publication criteria in order to increase the number of published county estimates.

3. Case Study: A Model for Planted Acres of Corn in Illinois

3.1 Past Research

The combination of multiple sources of data is a recurring theme throughout the NASS estimation program and at virtually all geographic levels. The traditional role of the ASB has been to assess several sources of information and come to a one-number, consensus estimate given that information. Model-based approaches to combining survey data with other sources of information have been pursued to help inform ASB yield estimates for select crops at state and regional levels (Wang et al. 2012, Nandram et al. 2014, Adrian 2012, Cruze 2015). Candidate models for combining multiple crop acreage data types were developed in Kim et al. (2016) to produce improved state indications.

NASS has actively pursued small area crop estimation although much of the research predates the full implementation of CAPS in 2011. Both unit-level models (see Walker and Sigman 1982, Battese et al. 1988, Stasny et al. 1991, Bellow 1993, Bellow and Lahiri 2012) and area-level approaches (Kott 1989; Bellow and Lahiri 2010, 2011, 2012; Williams 2013) have been investigated. Collectively, these efforts have touched on all four parameters of interest for major crops, including corn, soybeans and winter wheat. As shown in Table 1, the Battese-Fuller model output has been provided to the ASB as *an indication*. To date, no model-based estimate has served as the published official statistic.

3.2 An Area-Level Model for Planted Area

Following Fay and Herriot (1979), we define the following model for planted area totals within a state

$$\begin{aligned} Level\ 1: \hat{\theta}_i &= \theta_i + \epsilon_i && (3) \\ Level\ 2: \theta_i &= \mathbf{x}_i^{'}\boldsymbol{\beta} + u_i && (4) \end{aligned}$$

where $i \in \{1, ..., m\}$ is an index over m counties and $\hat{\theta}_i$ is the survey outcome for the i^{th} county computed from a realized sample of n_i positive reports of planted area in the county. The n_i could vary with crop type since a respondent may not grow all sampled commodities. Equation 3 describes the survey indication as an observation of the true planted area up to a sampling error ϵ_i, where it is assumed that $\epsilon_i \overset{indep.}{\sim} N(0, \hat{\sigma}^2_{\epsilon i})$, and the estimated sampling variances $\hat{\sigma}^2_{\epsilon i}$ can be obtained from the standard errors of the survey indications. The Level 2 model in Equation 4 describes the county-level total θ_i in terms of a linear function of some vector of observable covariates $\mathbf{x}_i$ and a county-specific random effect $u_i \overset{iid}{\sim} N(0, \sigma^2_u)$ assumed to be independent from the sampling errors ϵ_i.

Planted corn area in Illinois with reference year 2014 was selected as a test case because the state is a major producer of corn, the commodity is generally widely grown across the state, and a pool of covariates was readily available at the county level. For this case study, FSA administrative acreage for corn planted area was used at the county level. In addition, March precipitation provided by the National Oceanographic and Atmospheric Administration (NOAA) was incorporated since it could potentially delay or halt planting activity in the given year, thus, the vector of covariates for the i^{th} county consists of $\mathbf{x}_i = (FSA\ corn, NOAA\ March\ precip.)'$ in this case study.

Due to observed skewness in the distribution of the county survey indications $\hat{\theta}_i$, the decision was made to scale each indication by the number of positive reports n_i and model the transformed variable. Point estimates of the area totals obtained under Equation 3 and Equation 4 may not automatically sum to the published state totals. Benchmarking to the state total is necessary to ensure consistency of acreage and production estimates at state, ASD, and county levels.

3.2.1 External Benchmarking

In practice, NASS requires that official county and ASD estimates satisfy external benchmarks since state totals are published prior to the publication of any sub-state crop estimates. Therefore, model-based approaches that also honor Equation 1 and Equation 2 are desirable. Letting a denote the fixed and known state-level estimate, $n = \sum_{i=1}^{m} n_i$ denote the sum of all positive reports statewide, and $n_i\tilde{\theta}_i$ denote a modeled estimate transformed back to the scale of the i^{th} county total, a ratio benchmarking approach was considered. This non-parametric approach applies the same corrective factor shown in Equation 5 to each of the m modeled estimates $\tilde{\theta}_i$; after benchmarking, the m county-level totals agree with the established state total as in Equation 6.

$$\tilde{\theta}_i^{RB} = \tilde{\theta}_i * a \left(\sum_{k=1}^{m} n_k \tilde{\theta}_k \right)^{-1} \tag{5}$$

$$a = \sum_{i=1}^{m} n_i \tilde{\theta}_i^{RB} \tag{6}$$

This approach to benchmarking relates the county totals directly to the state total a. The models were formulated as Bayesian hierarchical models and fit by Markov chain Monte Carlo simulation using a program for Bayesian analysis. For brevity, the choice of prior distributions and other details of the algorithm are omitted here. Model estimates of the county totals and associated standard errors were obtained as posterior means and variances. Note that the ASD totals and standard errors may be obtained by summing corresponding iterates of the Markov chains of member counties. *This method of benchmarking automatically generates point estimates for counties and ASDs satisfying Equation 1 and Equation 2 while producing defensible measures of uncertainty for each.*

3.2.2 Preliminary Findings

Model-based estimates of area planted in corn for the 2014 crop year were computed for the 102 counties and 9 ASDs in the state of Illinois. Models were fit assuming no external benchmark (denoted ME) and incorporating the ratio benchmarking of the previous section (MERB). A comparison of coefficients of variation in Table 2 shows that either of the model-based strategies reduce the CV of the sub-state estimate relative to the survey direct expansion (DE) of planted area. In this case study, reductions in CV of nearly 50% and more were realized compared to the survey indication alone.

While offering reduction in CV, the model-based approach which does not incorporate the state total as an external benchmark (ME), does not have the desired accuracy. In Figure 5, the ME estimates are plotted against corresponding county- and ASD-level FSA administrative acreage data. Nearly all of the ME estimates fall below the 45 degree line indicating that they are smaller than the corresponding FSA administrative corn acreage indications. This highlights the list-based nature of the survey sample and illustrates that

Table 2: Coefficients of Variation (%) for Illinois corn planted area estimates in crop year 2014 for 102 counties within 9 Agricultural Statistics Districts

Level	Statistic	DE	ME	MERB
County	min	9.1	3.9	3.2
	median	19.2	6.9	6.8
	max	92.3	28.2	28.2
District	min	4.4	2.4	1.4
	median	6.8	2.7	1.8
	max	8.7	4.2	4.2

the act of benchmarking also plays a role in coverage adjustment for the list-only sample. Again, the FSA administrative data is generally interpreted as a lower bound on planting activity as it represents a minimum amount of planting known to have taken place within the county (or ASD) lines. A modeled estimate for planted area that fails to cover this administrative source of planted acreage may be viewed skeptically.

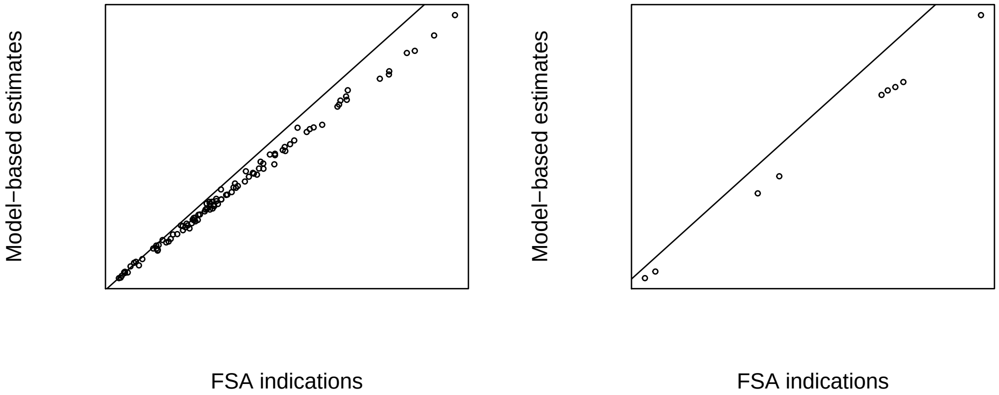


Figure 5: ME modeled county and ASD estimates of Illinois planted corn without benchmarking versus FSA administrative acreage data in the 2014 crop year

Figure 6 shows the resulting ratio benchmarked (MERB) estimates versus the same FSA data. All ASD-level estimates are strictly greater than the FSA data, however, even some MERB county estimates may fall below the FSA county-planted area totals for corn. Questions arise as to how the model-based estimate should be interpreted (either by the ASB during review or by end users of a model-based official statistic) with respect to FSA acreage data.

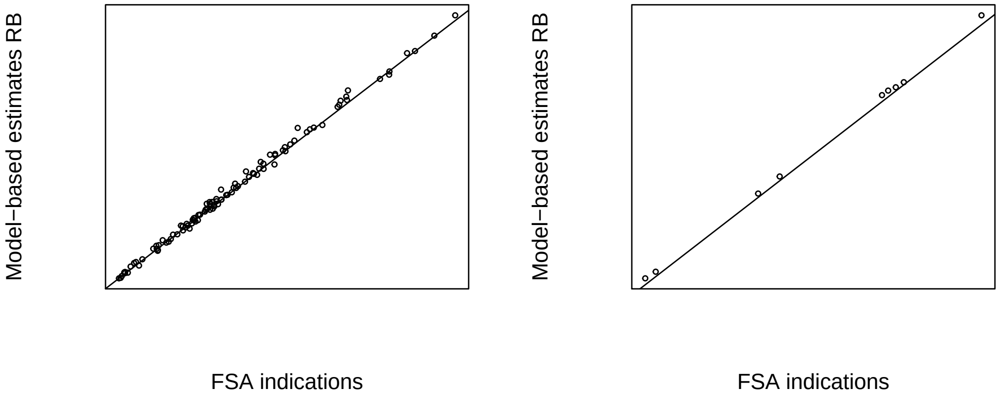


Figure 6: MERB modeled county and ASD estimates of Illinois planted corn incorporating ratio benchmarking versus FSA administrative acreage data in the 2014 crop year

4. Discussion and Conclusions

The NASS crops survey cycle is designed to capture crop activities from initial planting decisions to the final harvest and production. Starting in 2011, a larger probability-based sample for sub-state estimates was realized with the full implementation of CAPS. Even after strengthening the direct survey indications for counties and ASDs, issues of nonresponse, variation in planting decisions, and sparsity of some types of crops within county lines have led to the suppression of many NASS county estimates under the current publication standards.

Small area models have the potential to add value to the NASS crops county estimates program in terms of reproducibility and quantification of uncertainty while harnessing a wider variety of auxiliary data types. Whether the models are implemented as several separate indications in the ASB review process or ultimately vetted for use as the official statistic, developing a publication standard that can also utilize characteristics of models may be necessary in order to increase the number of published county estimates across the entire crops program.

Although the case study presented in this paper has focused on a model for planted area, the eventual goal is to support all parameters of interest with appropriate models. In particular, strengthening yield and production estimates may lead to a greater number of publishable counties. The proposed methodology can be extended to the other totals including harvested area and production. The extension to harvested area may be more straightforward as planted area and harvested area have a strong, positive correlation and some of the same administrative data for planted area could serve as useful covariates for harvested area. Some particular challenges exist for estimating the production totals as there is little in the way of rich county-level administrative data and item-level nonresponse tends to result in diminished sample sizes.

A few of the ongoing challenges are summarized below.

- *The best use of all available data*–At the authors' discretion, the model presented in this preliminary case study made use of just of one key auxiliary data source for acreage, namely the FSA planted area data. Both the FSA acreage and the Cropland Data Layer acreages are highly correlated. An important question worth investigating is whether the quality of the model has been affected by foregoing the use of the CDL data, and under what circumstances it might be a preferred source of auxiliary information.

Looking ahead to harvested area, production, and yield, NASS uses its survey data *twice* in its traditional estimation process: once to obtain a total indication, and once to construct a ratio indication. The merits of incorporating one or the other into a model should be carefully weighed. Benchmarking a ratio, e.g., yield, introduces additional challenges.

- *Structural zeros versus 'missing' indications*–In select cases, the survey may not indicate any planted area with respect a commodity, but some positive acreage may be represented by the FSA administrative data for the county. Due to item-level nonresponse associated with production, it is possible, in some cases, to have positive harvested acreage survey indications within a county but have no sample to support a production survey indication in that same county. This may be a particular problem in counties where a certain commodity crop is already sparse. The structural zero does not represent a problem for benchmarking to the state total; if a county did not contribute toward a state total, it can simply be omitted from the constraint. On the other hand, the 'missing' indication could potentially affect the quality of other county estimates through omission, therefore, an appropriate synthetic estimate for production may be desirable in such counties.

- *Implied relationships among estimates*–Related to the structural zero, a zero estimate for planted area implies a zero estimate for harvested area, and a zero estimate for harvested area implies a zero estimate for production. Furthermore, any defensible estimate of harvested area within a region must be no greater than the planted area estimate for that region. It is desirable for planted area estimates to at least cover FSA administrative acreages whenever possible. Since yield is a ratio of production to harvested area, once any two of the three point estimates are known, the third is determined. The model-based estimates obtained from several independent models may not automatically satisfy these relationships. NASS's traditional estimation and review process excels at enforcing these physical relationships in the official statistics. In part, this may be why NASS has preferred to use the Battese-Fuller acreage models as a separate source of planted acreage indications rather than the official statistic. Incorporating constraints into the models or jointly modeling county-level parameters may be worthwhile.

- *A meaningful pool of covariates for production*–In the absence of administrative data on production, other important characteristics including measures of soil productivity, weather, climate data, and remote sensing indices may be taken into consideration. Building the best pool of auxiliary data for the 43 CAPS states and the range of commodities to be supported by the crops county estimates program is a formidable task. Moreover, the best use of these data will require crop-specific knowledge about critical growing stages of each commodity and incorporating data at the appropriate temporal and spatial resolution. Absent administrative data, an appropriate 'gold standard' or best basis for comparison of models of yield and production is paramount.

NASS has engaged an expert panel under the auspices of the Committee on National Statistics (CNSTAT) to advise on the state of its county estimates program. The first two of six meetings took place in November 2015 and May 2016, with subsequent meetings to be held later in 2016 and in 2017. Themes for the 10 member expert panel have included NASS's current practices, timelines, and requirements for producing official county estimates; the NASS publication standard; users and uses of NASS county estimates data; record linkage of NASS, FSA, and remote sensing data sources; and potential additional sources of data, e.g., precision agriculture data. The panel is tasked with recommending strategies for incorporating models into the estimation program with the goal of increasing the number of published estimates and with identifying current and potential future data sources that could augment NASS's capabilities. The findings of the expert panel are to be released in a comprehensive report upon the conclusion of its sixth meeting.

While NASS awaits the recommendations and findings of the CNSTAT panel, research in small area estimation continues. Forthcoming work will investigate the applicability of sub-area level models to estimate planted and harvested crop area. Evaluation of these models for multiple states and for multiple commodities is underway. Modeling the yield ratio represents the additional challenge of benchmarking a weighted sum; candidate models for production and yield are under development and review. NASS will continue to define the role that model-based estimation should serve in the production of its official crops county estimates.

Acknowledgments

The authors wish to thank Ray Roberts and James Johanson of USDA NASS for producing Figure 4.

REFERENCES

Bailey J.T. and Kott, P.S. (1997), "An Application of Multiple List Frame Sampling for Multi-Purpose Surveys," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 496-500

Battese G.E., Harter R.M., Fuller W.A. (1988), "An Error Component Model for Prediction of County Crop Areas Using Survey and Satellite Data," *Journal of the American Statistical Association*, 83, 28-36.

Bellow M.E. (1993), "Application of Satellite Data to Crop Area Estimation at the County Level" *USDA NASS Report*, June.

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

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

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

Boryan C., Yang Z., Mueller R., Craig M. (2011) "Monitoring US Agriculture: the US Department of Agriculture, National Agricultural Statistics Service, Cropland Data Layer Program," Geocarto International 26 (5), 341-358

Congressional Budget Office (2014) Cost Estimate of the Agricultural Act of 2014 provided to the U.S. House of Representatives Committee on Agriculture [Accessed September 29, 2016] `https://www.cbo.gov/sites/default/files/113th-congress-2013-2014/costestimate/hr2642lucasltr00.pdf`

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.

Hurst B. (2015) "Big Data and Agriculture: Innovation and Implications" Statement of the American Farm Bureau Federation to the House Committee on Agriculture. October 28, 2015. [Accessed September 29, 2016] `http://agriculture.house.gov/uploadedfiles/10.28.15_hurst_testimony.pdf`

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

Johnson D. (2014) “An assessment of pre- and within-season remotely sensed variables for forecasting corn and soybean yields in the United States.” Remote Sensing of Environment 141,116-128.

Johnson D. (2016) ”A comprehensive assessment of the correlations between field crop yields and commonly used MODIS products”, International Journal of Applied Earth Observation and Geoinformation, 52, 6581.

Kim J.K., Wang Z., Zhu Z., Cruze N. (2016),“Combining Survey and Non-Survey Big Data for Improved Sub-Area Prediction Using a Multi-Level Model,” *Annals of Applied Statistics*, Submitted.

Kott P.S. (1989), “Robust Small Domain Estimation Using Random Effects Modeling,” *Survey Methodology*, 15, 1, 3-12.

Nandram B. and Sayit H. (2011), “A Bayesian Analysis of Small Area Probabilities Under a Constraint,” *Survey Methodology*, 37, 2, 137-152.

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

Stasny E.A., Goel P.K., Rumsey D.J. (1991) “County Estimates of Wheat Production”, Survey Methodology, 17, 211-225

Walker G., Sigman R. (1982) “The Use of LANDSAT for County Estimates of Crop Areas” NASS Technical Report. USDA NASS Education, Outreach: GIS Archived Reports. [Accessed September 29, 2016] `https://www.nass.usda.gov/Education_and_Outreach/Reports,_Presentations_and_Conferences/GIS_Reports/The%20Use%20of%20LANDSAT%20for%20County%20Estimates%20of%20Crop%20Areas%20Evaluation.pdf`

Williams, M. (2013) “Small Area Modeling of County Estimates for Corn and Soybean Yields in the U.S.” Presentation at *Federal Committee on Statistical Methodology Research Conference* `http://www.copafs.org/seminars/fcsm2013research.aspx`