Special Proceedings: ISBN #: 978-81-950383-2-9
25th Annual Conference, 15-17 February 2023; pp 51-78

# Overcoming Challenges Associated with Early Bayesian State Estimation of Planted Acres in the United States

**Balgobin Nandram**
*Department of Mathematical Sciences*
*Worcester Polytechnic Institute, Worcester, MA, 01609, USA*

Received: 15 June 2023; Revised: 22 July 2023; Accepted: 24 July 2023

**Abstract**

National surveys in the United States have become expensive with low response rates, and there is an abundance of administrative data (non-probability samples). Government agencies are now beginning to integrate these two sources of data to improve the quality of official statistics. Our application is on agriculture, where the study variable is planted acres and estimates early in the current year are much needed by the USDA's National Agricultural Statistics Service (NASS). A solution of this problem is important for economic, policy and many other reasons. This is a very difficult problem to solve because there are many challenges, including the poor quality of the available early survey data, that must be overcome. We attempt to solve the problem by integrating the probability samples from designed surveys and the non-probability samples, relatively much larger, which come from 'administrative' data or 'historical' data. Keeping in line with NASS's preference, we use Bayesian small area temporal models (a non-spatial model and a spatial model) to infer the early state estimates of planted acres. The Bayesian Fay-Herriot model is manipulated to link the data, and the Gibbs sampler, which is operationalized, is used to fit the two models. We show that the spatio-temporal model provides higher quality state estimates than the non-spatio-temporal model.

*Key words*: Conditional autoregressive (CAR) model; Data integration and data quality; Fay-Herriot model; Gibbs sampler; Non-probability samples; Structural error models.

**AMS Subject Classifications**: 62F15, 62D05, 62D10, 62P12

## 1. Introduction

It is the objective of this paper to show how to estimate planted acres for states early in the season in the United States. These estimates are based on historical data, administrative data and survey data. Estimates of planted acres are so important that for farmers and price analysts, almost every discussion of crop fundamentals begins with planted

Corresponding Author: Balgobin Nandram; Email: balnan@wpi.edu, balgobin.nandram@usda.gov
This paper is based upon the Keynote Address I delivered on 15 February 2023 during the 25th Annual International Conference of the SSCA held at University of Jammu during 15-17 February 2023.

acres (*e.g.*, Kansas Farm Bureau, 2020). It is our purpose to demonstrate how to integrate these data sources with limited access to the actual survey data, which are confidential, and instead we used published data. The key problem is to provide estimates of planted acres in June of any year, when the quality of the actual survey data is poor, and one needs to access other data, most of which are available to the public. We imagine that the current year is 2021 and estimates of planted acres are required in June. The methodology is being developed so that it can be used readily for June of any year just after the survey data are available. However, it is not the purpose of this paper to present methodology or substantial results from modeling real data; rather it is intended to show critical thinking on, and the struggles in the original stages of, this project at the National Agricultural Statistics Service.

We have Farm Service Agency (FSA) data, which are administrative data. These are voluntary to the farmers and are essentially a non-probability sample. In our models, we integrate survey data and non-survey data in the "current year" (2021), collected in June, and all available data over the past decade. We work at state level because county level data are not available in June. We include all possible data sources. It is required that all model estimates must cover (*i.e.*, larger than) FSA planted acres. We also have final results from the Agricultural Statistics Board (ASB). Both FSA and ASB values are historical data (before 2021). FSA values are not available in June of the current year but ASB values are available in March and June; see The National Agricultural Statistics Service (2021 a) for prospective plantings in March and The National Agricultural Statistics Service (2021 b) for acreage in June. We have analyzed planted acres (thousands of acres) for corn, which is our focus, and a similar analysis can be done for other crops such as soybeans.

NASS conducts quarterly Agricultural Production Survey (APS) in an ongoing effort to capture activities throughout the life cycle of the crop. These include planting intentions (March), early estimates of planted acreage (June, with some intentions), and output activities for small grains crops (September) such as buckwheat, flax, oats and rye, and major row crops (December) such as corn, soybeans, cotton, potato. The June Area Survey (JAS) provides an under-coverage adjustment for the list-based samples obtained during the June, September and December APS surveys.

According to Young and Chen (2022), "The NASS conducts more than a hundred national surveys and produces more than 400 reports each year. An annual publication calendar details the day and time each report is to be released, and the NASS has consistently released its reports according to schedule more than 99% of the time." The NASS acreage and production reports are considered by many to be the "final word" because they are **Unbiased** (they are not influenced by either buyers or sellers of commodities); **Timely** (data are provided well in advance of when they will be available from other sources); **Consistent** (the same statistically sound procedures are followed each time, building on a multi-year data-base); and **Transparent** (NASS ensures that all participants have equal access to the information). For further discussions of these notions, see, for example, Kansas Farm Bureau (2020). The Research Development Division (RDD) at NASS is charged to ensure that all procedures are current, and if not, they are revised and new methods are developed.

There are several reasons why early estimates of planted acres are needed in the

United States.

1. June Acreage Report is a very important economic indicator in the United States and the amount of planted acres affect prices later. It follows the well known demand and supply principle in economics.

2. Stake holders and economists need high quality estimates, and it is incumbent on NASS to provide these estimates. NASS is required by law (Agricultural Marketing Act of 1946 and the Census of Agriculture Act of 1997) to produce estimates for several key crops as early as March. Markets are hungry for information.

3. Budgets are allocated to different programs around this time.

4. Many internal programs at NASS depend on quality estimates of planted acres (*e.g.*, cash rental rates), so these numbers must be reported early.

5. Even before March, many farmers are, or are considering marketing portions of their expected production.

6. To make the most informed decision, farmers, agribusinesses and even speculators need as accurate a picture as possible as what market fundamentals are and how those fundamentals are changing as the year progresses. USDA's NASS provides objective information to all market participants at the the same time at no cost.

Now, we give an idea of the order of magnitude of the APS and the JAS. March APS has about 80,000 US farm operators, a survey of farmers conducted in the first two weeks of March to get intentions, selected from a list of farmers that ensure that all operations had a chance to be selected. Note that intentions are not binding, and the farmers could change their minds, and this is a difficulty that is impossible to address in June. Like all NASS surveys, data are collected by mail, internet, telephone, and personal interviews. June Crop Acreage report, which includes two surveys, the APS, a survey of over 70,000 farmers, asking the farmers how many acres they had planted and still intended to plant, and the JAS, which includes over 11,000 individual (one square mile) segments, in which enumerators physically inspect, to see what has been planted (and then ask the farmers what will be planted on any unplanted tracts in the segment). This is a dual-frame survey and the two surveys are combined to complete the June Crop Area estimates.

Every farmer participating in the USDA Farm Service Agency (FSA) programs, such as marketing assistance loans or deficiency payments, must file an FSA-578 Report of Acreage. However, the acreage report deadline is July 15 for FSA (not March or June), and not every farmer gets it on time, and not every county office office gets the data inputted immediately. Also, not every farmer participates in the FSA programs. Consequently, the August FSA reports underestimates planted acres. Therefore, it is still an important constraint that must be incorporated into our model for the June estimates of planted acres; see Office of the Chief Economist (2019). It turns that it is a difficult problem to incorporate the constraint directly into the model, but this is not our purpose in this paper; see Nandram

*et al.* (2021, 2023) and Chen, Nandram and Cruze (2022) of work already done at NASS, where the constraints are placed directly into the models.

Next, we discuss the challenges that we must overcome to provide estimates of planted acres with reasonable and satisfactory quality. It is pertinent to list the challenges here.

1. FSA current year values are not available in June. They become available later in August.
2. The dual-frame (APS/JAS) model estimates must be larger than the FSA values.
3. State survey indications do not capture variation very well.
4. There are outlying states (some very large and some very small).
5. With the initialization of modernization and unification at NASS, we want to combine administrative data (non-probability sample) with the surveys (APS/JAS). Historical data (available) are incorporated as the non-probability sample; we have 10 years of ABS/FSA data before the current year.
6. NASS wants model estimates for 48 US states (excluding Alaska and Hawaii). Typically data for corn may be available for all 48 states with missing survey data; soybeans are available from fewer number of states, actually 29.
7. Weather variables (temperature and precipitation) are difficult to use, although they are important. Current work at NASS is now trying to make use of the weather variables.
8. Landsat satellite (imagery) data are of poor quality in June, and they are not useful; in March there are only intentions.
9. Covariates must be incorporated as well; there are missing values here also.
10. Meeting the annual production schedule is difficult.
11. National Academy of Sciences, Engineering and Medicine (2017) recommended that NASS use Bayesian Small Area models. These models are complicated, and Markov chain Monte Carlo methods (ıe.g., Gibbs samplers) are needed to fit them.

A non-probability sample and a probability sample can be combined in several ways. This depends on available data; see Rao (2020) for both design-based and model-based approaches for making valid inferences by integrating data from surveys and other sources. Also, Li, Chen and Wu (2020) presented double robustness with quasi-randomization via propensity scores. Nandram, Choi and Liu (2021) and Nandram and Rao (2021, 2023) provided Bayesian analyzes. But these can be carried out when survey weights are available from the probability sample. In the current work, survey weights are already incorporated into the survey indications for states, and combining the two samples need an alternative approach. We use a measurement error model to combine the two samples; see Fuller (1987)

and Berg *et al.* (2021). In our model, there is a linear relation between the FSA values and the ASB values for the historical data (the non-probability sample), and this same relation holds between the current year's FSA values and the true value of planted acres. This permits integration of the two data sources.

This paper has five sections including the current one, Section 1. In Section 2, we describe the available data. In particular, we describe how to estimate the FSA values before June of the current year, 2021. We also show how to impute the missing indications and variances. In Section 3, we describe the temporal models, a non-spatial model and a spatial model, which we use for comparison. We also describe the computations, and model diagnostics. In Section 4, we present the data analysis of the public-used data. Section 5 has concluding remarks. Appendix A contains a short description on how to go down to the level of Agricultural Statistics Districts (ASD) for further analysis. Appendix B has a brief description of how to deal with clustering in the indications. Appendix C contains a list of abbreviations used in the paper.

## 2. Available data and FSA values

In this section, we give a more detailed discussion of the data we must use to exemplify the actual situation. We primarily study corn, but there are other crops of interest such as soybeans, All wheat and All cotton; again see The National Agricultural Statistics Service (2021 a) for prospective plantings in March and The National Agricultural Statistics Service (2021 b) for acreage in June.

### 2.1. General data

We have Farm Service Agency (FSA) and Agricultural Statistics Board (ASB) historical data for the past ten years before 2021, and these are not confidential. Our idea is that the relation between the FSA values and the ASB values should be similar to the relation between the FSA values and the true planted acres in the current year. This is how the non-probability sample (FSA values and ASB data) are used. In March, there are indications on planting intentions, approved by the ASB; the March ASB values are also available for the past ten years. As stated, we have 10 years of FSA values before 2021, but not in 2021, which we need. Note that ASB values are available to the public. We have the Agricultural Production Survey (APS) and June Area Survey (JAS) dual frame survey indications, but these are confidential, they are not available for the public use, and they are not used in this paper. However, approved estimates are available in June for the public, and these are the ones used in this paper for exploratory analysis. There is on-going work on the actual data at NASS.

We also have five covariates, which are Percent farmland irrigated - $x_2$, Population density - $x_3$, Value of cropland - $x_4$, National commodity crop production index (NCCPI), an index of soil quality, - $x_5$, Number of farms - $x_6$. These are publicly available. A simple regression of June 2021 survey indications on the covariates gives an $R^2 \approx 75\%$; $x_4$ and $x_6$ are significant; $x_2$, $x_3$, $x_5$ and $x_4 * x_6$ are not significant. Other variables such as weather (temperature and precipitation) are currently being explored at NASS.

In Figure 1, we show the maps of quintiles of the FSA values and the survey indications of planted acres in the current year. We can see some differences (e.g., OK moves from 3 to 2, OH from 4 to 5, AL from 2 to 3).

## 2.2. Imputing missing data

We show how to impute the FSA values of the current year (2021). Then we show how to impute missing indications and variances. The public use data on planted acres do not come with estimated variances, which are needed in the Fay-Herriot model; see The National Agricultural Statistics Service (2021 a, b).

### 2.2.1. Current year FSA values

We use $T+1=10$ (*i.e.*, $T=9$) years of FSA values and March intentions (put out by ASB) to predict the current year FSA values. We denote the ten years of historical data by

$$(\hat{\theta}_{it}^{(f)}, \hat{\theta}_{it}^{(a)}), i=1,\ldots,\ell, t=-T,\ldots,0.$$

Note that $t=1$ is the current year, the year of interest. Then, we use simple linear regression,

$$\hat{\theta}_{it}^{(f)} = \beta_0 + \beta_1 \hat{\theta}_{it}^{(a)} + e_i, i=1,\ldots,\ell, t=-T,\ldots,0.$$

We fit this model to get the following estimates of the regression coefficients. The 10 speculative states for corn gave $\hat{\beta}_0 = -58.05, \hat{\beta}_1 = .978, R^2 \approx 1$. Although it is not particularly relevant, the 11 speculative states for soybeans gave $\hat{\beta}_0 = 13.03, \hat{\beta}_1 = .986, R^2 \approx 1$. Therefore, the fits are pretty good for both corn and soybeans. However, there are some aberrations for smaller states (38 for corn and 18 for soybeans). Finally, we predict the current year FSA values from

$$\hat{\theta}_{1i} \equiv \hat{\theta}_{i1}^{(f)} = \hat{\beta}_0 + \hat{\beta}_1 \hat{\theta}_{i1}^{(a)}, i=1,\ldots,\ell.$$

The $\hat{\theta}_{1i}$ will be used as part of the observed data in this paper or at NASS.

It is possible to improve this procedure using covariates such as precipitation and temperature (under study at NASS).

### 2.2.2. Missing indications and variances

Fewer than 48 states are observed for corn and fewer than 29 states for soybeans; some states are missing both indications and variances. We use the adjacent neighbors of a specific state without indications and/or variances via an incidence matrix to impute the remaining states for corn. The same can be done for soybeans (currently under experimentation at NASS) and other crops such as All wheat and All cotton.

Let $\mathcal{C}_i$ denote the set of adjacent neighbors of the $i^{th}$ state, and let $n_i$ denote the number of counties in the $i^{th}$ state. Then, if the $i^{th}$ state's indication and/or variances are missing, define

$$\hat{\theta}_i = \frac{\sum_{j\in\mathcal{C}_i} n_j \hat{\theta}_j}{\sum_{j\in\mathcal{C}_i} n_j} \quad \text{and} \quad \hat{\sigma_i^2} = \left\{ \prod_{j\in\mathcal{C}_i} (\hat{\sigma_j^2})^{n_j} \right\}^{\frac{1}{\sum_{j\in\mathcal{C}_i} n_j}},$$

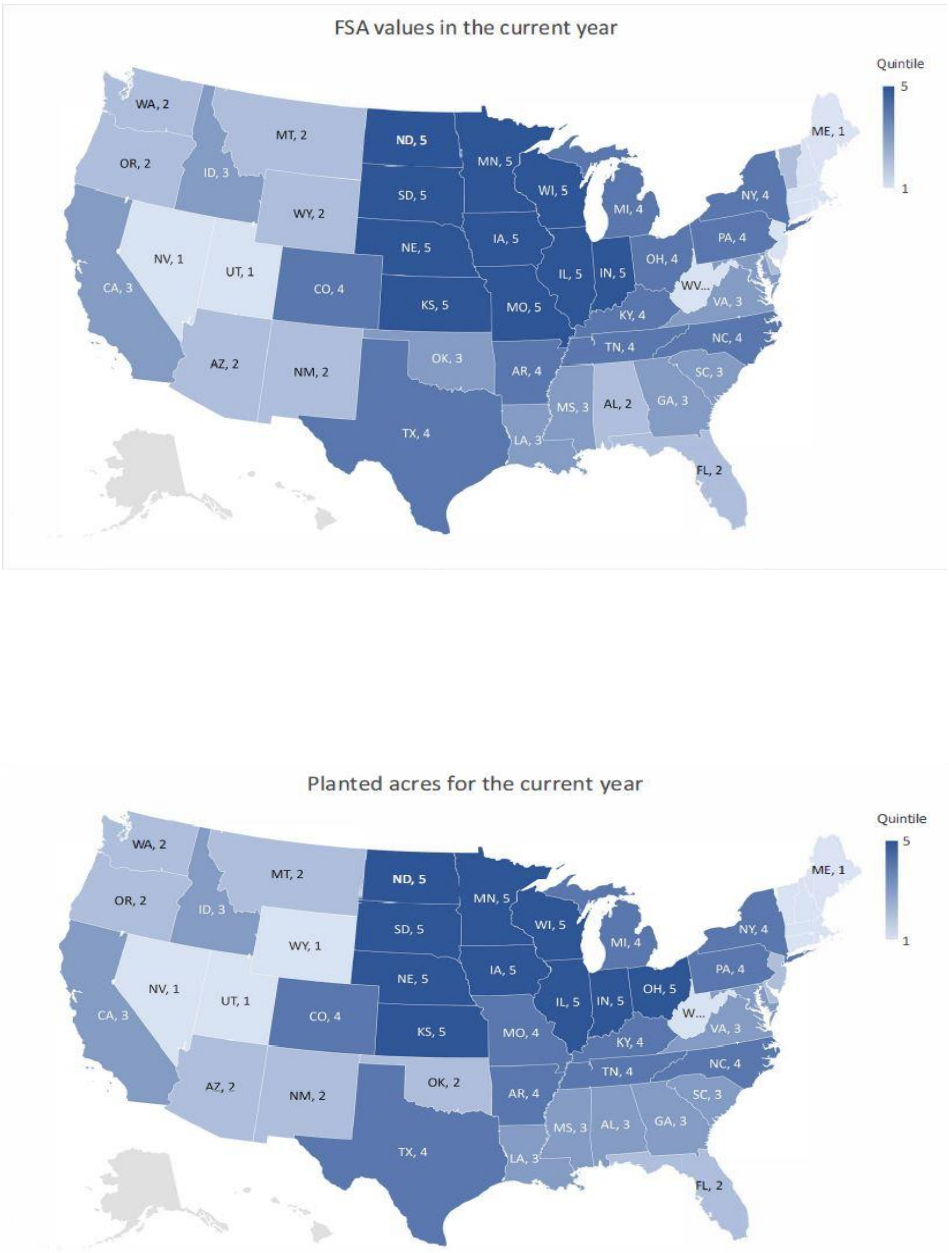


**Figure 1: Quintiles of FSA values and survey indications of planted acres: The quintiles for the FSA values (survey indications) are** 74 (85), 293 (330), 597 (640), 3276 (3350), **and for the FSA values (survey indications), the minimum and maximum values are** 1.89 (2.00) **and** 12323 (13100)**.**

the weighted arithmetic mean and the weighted geometric mean for respectively the indications and variances. These are the 'observed' indications and variances corresponding to the missing states in the public-used data.

Public-used ASB indications do not come with variances, but these are confidential data available at NASS for the survey indications. Because of confidentiality, we cannot use these data in this paper. Besides the variances are too small (optimistic) because of the large amount of data that go into a state indication. When the state indications are obtained, heterogeneity and clustering in the data are not taken into account. The data are weighted (to reflect the survey design) from the operation (farm) level to state level.

Assuming that $\hat{\theta}_i$ is observed, we take

$$\hat{\sigma}_i^2 = CV_i^2 \times \hat{\theta}_i^2, i = 1, \ldots, \ell,$$

where $CV_i$ is the coefficient of variation and the $\hat{\theta}_i$ are the state indications. Here $CV_i$ is also unknown, so we take

$$CV_i = \text{Uniform}(.10, .50), i = 1, \ldots, \ell,$$

because a coefficient of variation of .30 is taken to be a threshold at most government agencies in the United States. Here $\ell = 48$ for corn and 29 for soybeans. This procedure is a bit problematic because it penalizes some large states and some small states appear too good.

An alternative and slightly better procedure is to take $CV_i$ to be inversely proportion to the number of counties, $n_i$, in the $i^{th}$ state, and $\frac{1}{\ell}\sum_{i=1}^{\ell} CV_i = .30$, again a threshold for a reliable estimate in US government agencies. This gives

$$CV_i = \max\left(.10, \frac{0.30}{\frac{1}{\ell}\sum_{i=1}^{\ell}\frac{1}{n_i}}\right), i = 1, \ldots, \ell,$$

where $n_i$ is the number of counties in the $i^{th}$ state, and for flexibility, the $CV_i$ can be kept larger than 0.10 for speculative states. But we have not done so for this paper. It is another difficult problem to specify the coefficient of variations, and clearly more data are needed.

As a summary, we present the data we want to analyze. The true values that we want to estimate are denoted by $\underline{\theta} = (\theta_i, i = 1, \ldots, \ell)$, where $\ell$ denotes the number of states. This will vary with different commodities, but as was stated we will deal only with planted acres (thousands of acres) for corn. We denote the data by $D$, where

$$D = \{\underline{\hat{\theta}}^{(f)}, \underline{\hat{\theta}}^{(a)}, \underline{\hat{\theta}}_1, \underline{\hat{\theta}}_2, \underline{\hat{\sigma}}_2^2\}.$$

The (FSA, ASB) historical values are $(\hat{\theta}_{ti}^{(f)}, \hat{\theta}_{ti}^{(a)}), t = -T, \ldots, 0$ ; the current year FSA values are $\hat{\theta}_{1i}$, obtained by imputation; the current year survey indications and variances are $(\hat{\theta}_{2i}, \hat{\sigma}_{2i}^2)$ in June, obtained from the APS and the JAS; and the covariates are $\underline{x}_i, i = 1, \ldots, \ell, c = 6$, including an intercept. We adapt the Fay-Herriot model, and a novel simpli-

fication is to introduce the ratios,

$$\kappa_i = \sigma_0^2/\hat{\sigma}_i^2, i = 1, \ldots, \ell, \sigma_0^2 = \text{ GM of } \hat{\sigma}_i^2,$$

where GM stands for geometric mean, and the $\kappa_i$ are assumed known. We will introduce two models, which use all the available data, and they are a non-spatio-temporal (NST) model and spatio-temporal (ST) model; the $\theta_i$ are linked to the covariates, $\underline{x}_i$, via regression with unknown coefficients. There are two additional features: First the models take care of outliers, and second, the constraint, $\theta_i > \hat{\theta}_i, i = 1, \ldots, \ell$, is not part of the models, but it is taken care of in the output analysis. This avoids model complications. For a single model, Bayesian diagnostics are not appropriate if the constraint is included because they check how close the predictive data are to the observed data.

## 3. Bayesian small area models

Small area models are appropriate because the indications from many small states are not reliable. For survey data, there is only one data point for each state. There are also supplemental data for ten years before the current year, and the FSA imputed value for each state of the current year, 2021. In Section 3.1, we present the two models and we briefly describe the computation. In Section 3.2, we present a diagnostic assessment of the two models.

### 3.1. Models and computations

We first describe the non-spatial model. The spatial model is similar except with one adjustment.

For the historical data, we assume

$$\hat{\theta}_{ti}^{(f)} \mid \{\hat{\theta}_{ti}^{(a)}, \alpha_0, \alpha_1, \psi_1, \sigma^2\} \overset{ind}{\sim} \text{Normal}(\alpha_0 + \alpha_1 \hat{\theta}_{ti}^{(a)}, \psi_1 \sigma^2), \tag{1}$$

$t = -T, \ldots, 0, i = 1, \ldots, \ell$, and for the current year's FSA values, we assume

$$\hat{\theta}_{1i} \mid \{\alpha_0, \alpha_1, \theta_i, \sigma^2\} \overset{ind}{\sim} \text{Normal}(\alpha_0 + \alpha_1 \theta_i, \sigma^2). \tag{2}$$

For indications and variances, we assume

$$\hat{\theta}_{2i} \mid \{\theta_i, z_i = 0, p, \sigma^2, \psi_2\} \overset{ind}{\sim} \text{Normal}(\theta_i, \psi_2 \frac{\sigma^2}{\kappa_i}),$$

$$\hat{\theta}_{2i} \mid \{\theta_i, z_i = 1, p, \sigma^2, \psi_2\} \overset{ind}{\sim} \text{Normal}(\theta_i, \frac{\sigma^2}{\kappa_i}) \tag{3}$$

$$z_i \mid p \overset{ind}{\sim} \text{Bernoulli}(p)$$

$$\theta_i \mid \{\underline{\beta}, \sigma^2, \rho\} \overset{ind}{\sim} \text{Normal}(\underline{x}_i' \underline{\beta}, \frac{\rho}{1-\rho} \sigma^2), \tag{4}$$

and

$$\pi(\alpha_0, \alpha_1, p, \rho, \psi_1, \psi_2, \underline{\beta}, \sigma^2) \propto \frac{1}{\sigma^2}\text{Beta}(\frac{\sqrt{\ell}}{2}, \frac{\sqrt{\ell}}{2}). \tag{5}$$

The beta prior is used for stability, and it is motivated by minimum mean square error; see Casella and Berger (2002, *pg.* 332) for example. The mixture model, used to accommodate outliers and robustness to normality, is an extension of the Fay-Herriot model (Fay and Herriot, 1979). Also, see Goyal, Datta and Mandal (2020) for a slightly different formulation of the mixture model.

Because the parameters are weakly identified in the survey part of the model, there is a need to specify bounds for $\alpha_1$ and $\alpha_2$, and we do so using an exploratory data analysis, namely $a_0 < \alpha_1 < a_1$, $b_0 < \alpha_2 < b_1$. We also specify $c_0 < \rho < c_1$. We believe the relation in (1) is tight so we assume $0 < \psi_1 < 1$. We also assume that $0 < \psi_2 < 1$ because outliers should be more variable than non-outliers, and $0 < p < 1/2$ because there should be fewer outliers than non-outliers. These latter assumptions are natural, and all constraints are incorporated into the model when it is fit using the Gibbs sampler. However, for simplicity, the constraint, $\theta_i > \hat{\theta}_{1i}, i = 1, \ldots, \ell$, that the model estimates are larger than FSA values is incorporated into the output analysis, not within the Gibbs sampler.

Note that the non-probability sample and the probability sample are linked by (2), and (1) and (2) have the same regression coefficients.

For the spatial model, we use the conditional auto-regressive (CAR) model,

$$\underline{\theta} \mid \{\underline{\beta}, \sigma^2, \rho\} \overset{ind}{\sim} \text{Normal}\{X\underline{\beta}, \frac{\rho}{1-\rho}\sigma^2(R - \psi_3 W)^{-1}\}, X = (\underline{x}_i'), \tag{6}$$

where $\lambda_1, \ldots, \lambda_\ell$ are eigenvalues of $R^{-1}W$ in increasing order (some negative and some positive). We simply replace (4) by (6) with an extra parameter, $\psi_3$, beyond the less flexible intrinsic ($\psi_3 = 1$) CAR model (Janicki *et al.* 2022). A priori, we assume

$$\pi(\alpha_1, \alpha_2, p, \rho, \psi_1, \psi_2, \psi_3, \underline{\beta}, \sigma^2) \propto \frac{1}{\sigma^2}\text{Beta}(\frac{\sqrt{\ell}}{2}, \frac{\sqrt{\ell}}{2}), 0 < \psi_1, \psi_2 < 1, \frac{1}{\lambda_1} < \psi_3 < \frac{1}{\lambda_\ell}, \tag{7}$$

replacing (5) by (7). The NST model and the ST model are discussed in great detail in Nandram (2022), but this report is confidential. An earlier discussion is given by Berg et al. (2021); many issues in that paper are addressed in the report. This is part of the general measurement error model (*e.g.*, Fuller, 1987).

Let $\Omega = (\alpha_1, \alpha_2, \underline{\beta}, \psi_1, \psi_2)$ for the non-spatial model, $\Omega = (\alpha_1, \alpha_2, \underline{\beta}, \psi_1, \psi_2, \psi_3)$ for the spatial model ($\psi_3$ is not in the nonspatial model), and $D = \{\underline{\theta}^{\widehat{(f)}}, \underline{\theta}^{\widehat{(a)}}, \underline{\hat{\theta}}_2, \underline{\hat{\theta}}_2, \underline{\hat{\sigma^2}}\}$ denote the data. Then, using Bayes' theorem, the joint posterior density is

$$\pi(\Omega, \underline{z}, p, \underline{\beta}, \underline{\theta}, \sigma^2 \mid D).$$

We state the following steps in the griddy Gibbs sampler.

1. Integrate out $\underline{\theta}$ to get
$$\pi(\Omega, \underline{z}, p, \underline{\beta}, \sigma^2 \mid D).$$

2. Draw $(\underline{z}, p)$ together (collapsing and blocking),
$$\pi(\underline{z}, p \mid \Omega, \underline{\beta}, \sigma^2, D) = \pi(\underline{z} \mid p, \Omega, \underline{\beta}, \sigma^2 \mid D)\pi(p \mid \Omega, \underline{\beta}, \sigma^2 \mid D).$$

3. Draw $(\underline{\beta}, \sigma^2)$ together (collapsing and blocking),
$$\pi(\underline{\beta}, \sigma^2 \mid \Omega, \underline{z}, p, D) = \pi(\sigma^2 \mid \Omega, \underline{z}, p, \underline{d})\pi(\underline{\beta}, \mid \Omega, \underline{z}, p, \sigma^2, D).$$

4. Sample $\pi(\Omega \mid \underline{z}, p, \underline{\beta}, \sigma^2, D)$.

5. Monitor convergence (Geweke test and effective sample size).

6. Sample the Rao-Blackwellized density, $\pi(\underline{\theta} \mid \Omega, \underline{z}, p, \underline{\beta}, \sigma^2, D)$, in the output analysis subject to constraints (model estimates must cover FSA values). These are truncated univariate normal densities for non-spatial model and truncated multivariate normal densities for spatial model.

Markov chain Monte Carlo methods (Gibbs sampler with some collapsing and blocking to improve convergence and better mixing) are used to fit the two models; see Liu (1994) for collapsing and Tan and Hobert (2009) for blocking. In fact, we use the griddy Gibbs sampler (Ritter and Tanner, 1992) in which some CPDs are sampled using the grid method. The constraints are not included in the models to allow them to be as simple as possible, rather they are performed in an output analysis. In the non-spatial model, this is straight forward as we can sample from independent truncated normal densities, but in the spatial model, we need to sample from truncated multivariate normal densities (Ridgeway 2016). The constraint $\theta_i > \hat{\theta}_i, i = 1, \ldots, \ell$, in the output analysis.

In Table 1 we show the good performance of the Gibbs sampler under both models. Specifically, the Geweke tests show that the Gibbs sampler is stationary with all p-values being larger than .05 and the effective sample size (ESS) of each parameter is the nominal value of 1000, except the one for $\rho$ under the ST model, but this is still good. This shows that the two Gibbs samplers are strongly mixing. Also, note that the computational times are also operational at NASS; see the note to Table 1.

### 3.2. Model diagnostics

We use standard Bayesian diagnostics to check the goodness of fit of the two models. We assess the more interesting mixture part of the model (*i.e.*, the survey data).

We start by computing two simple diagnostic measures. Let $PM_i$ and $PSD_i, i = 1, \ldots, \ell$, denote the posterior means and posterior standard deviations from the two models.

Specifically, we have computed

$$ARES = \sqrt{\frac{1}{\ell}\sum_{i=1}^{\ell}(\hat{\theta}_{2i} - PM_i)^2}, \quad ASTD = \sqrt{\frac{1}{\ell}\sum_{i=1}^{\ell} PSD_i^2}.$$

For the non-spatial (spatial) model, we have $ARES = 2388$ (2005) and $ASTD = 366$ (120), showing the spatial model has performed much better than the non-spatial model in terms of these two measures. It is very good for the spatial model that it provides estimates closer to the direct estimates (indications) with smaller posterior standard deviations.

As a further check on the models, we have done a Bayesian cross-validation analysis (*i.e.*, delete one observation and predict it). The idea is the same for both models, but the specific mathematical formulas are different for the non-spatial model and the spatial model. Define

$$f(\hat{\theta}_{2i} \mid \underline{\hat{\theta}}_{(2i)}) = \sum_{h=1}^{M} W_{ih} f(\hat{\theta}_{2i} \mid \underline{\hat{\theta}}_{(2i)}, \Omega^{(h)}), W_{ih} = \frac{\{f(\hat{\theta}_{2i} \mid \Omega^{(h)}\}^{-1}}{\sum_{h=1}^{M}\{f(\hat{\theta}_{2i} \mid \Omega^{(h)}\}^{-1}}, i = 1, \ldots, \ell.$$

The residuals are $r_i = \hat{\theta}_{2i} - \mathrm{E}(\theta_{2i} \mid \underline{\hat{\theta}}_{(2i)}), i = 1, \ldots, \ell$. Then, a dispersion measure (DM, Wang *et al.* 2011), which we have developed, is

$$DM_1 = \frac{1}{\ell}\sum_{i=1}^{\ell} |r_i|,$$

and as this measure is not invariant to scale, we have now modified it to

$$DM_2 = \frac{1}{\ell}\sum_{i=1}^{\ell} \frac{|r_i|}{\mathrm{Std}(\theta_{2i} \mid \underline{\hat{\theta}}_{(2i)})}.$$

We also counted the number, $n_0$ of $r_i > 0$, the number, $n_3$, of $|r_i| \geq 3$ and the number, $n_4$, of $|r_i| \geq 4$. For the non-spatial (spatial) model, we got $DM_1 = 1144$ (110), $DM_2 = 5.61$ (0.91), $n_0 = 28$ (22), $n_3 = 29$ (13), $n_4 = 22$ (8). The spatial model is much better than the non-spatial model under these measures.

We have also calculated three standard Bayesian diagnostics with respect to the survey indications, $\hat{\theta}_{2i}$, which are the deviance information criterion (DIC), the Bayesian predictive p-value (BPP) and the log-pseudo marginal likelihood (LPML). The DICs are 875 (803), the BPPs are .399 (.594) and the LPMLs are $-417$ ($-419$) for the non-spatial (spatial) model. For the BPP and LPML there is basically no preference. However, the DIC does show that the spatial model is significantly better than the non-spatial model.

Finally, we compute the average absolute relative deviation (AARD) and the square root of the average squared relative deviation (RASRD), where we compare the posterior

**Table 1: Gibbs sampler diagnostics (*p*-values of Geweke test and effective sample sizes)**

| | Non-spatial | | Spatial | |
|---|---|---|---|---|
| | *P-val* | *ESS* | *P-val* | *ESS* |
| $\beta_1$ | .80 | 1000 | .67 | 1000 |
| $\beta_2$ | .77 | 1000 | .25 | 1000 |
| $\beta_3$ | .29 | 1000 | .63 | 1000 |
| $\beta_4$ | .38 | 1000 | .84 | 1000 |
| $\beta_5$ | .48 | 1000 | .66 | 1000 |
| $\beta_6$ | .92 | 1000 | .83 | 1000 |
| $\sigma^2$ | .19 | 1000 | .74 | 1000 |
| $\alpha_1$ | .21 | 1000 | .60 | 1000 |
| $\alpha_2$ | .22 | 1000 | .47 | 1000 |
| $p$ | .75 | 1000 | .06 | 1000 |
| $z$ | .85 | 1000 | .60 | 1000 |
| $\rho$ | .44 | 1000 | .97 | 884 |
| $\psi_1$ | .18 | 1000 | .63 | 1000 |
| $\psi_2$ | .61 | 1000 | .19 | 1000 |
| $\psi_3$ | – | – | .09 | 1000 |

NOTE: For the non-spatial model, the Gibbs sampler is run $55,000$ times, with a "burn in" of $5,000$ and we pick every $50^{th}$ one and this takes 3 minutes; for the spatial model, the Gibbs sampler is run $75,000$ times, with a "burn in" of $15,000$ and we pick every $60^{th}$ one and this takes 49 minutes. Here $z$ is the number of outliers.

means of planted acres to last years ASB values as

$$AARD = \frac{1}{\ell}\sum_{i=1}^{\ell}\frac{|PM_i - ASB_i|}{ASB_i}, \quad RASRD = \sqrt{\frac{1}{\ell}\sum_{i=1}^{\ell}\left\{\frac{PM_i - ASB_i}{ASB_i}\right\}^2}.$$

We expect the current year's ASB values, which are unknown, to be similar to last year's. In Table 2 we show that the ST model does better than the NST model; the numbers under the ST model are smaller than those under the NST model. Specifically, the spatio-temporal (ST) model has smaller AARD and RASRD values than under the non-spatio-temporal (NST) model with or without the constraints.

**Table 2: Average absolute (squared) relative deviation by model and constraint**

| Constraint | *AARD* | | *RASRD* | |
|---|---|---|---|---|
| | *NST* | *ST* | *NST* | *ST* |
| No | 0.240 | 0.209 | 0.349 | 0.315 |
| Yes | 0.272 | 0.229 | 0.391 | 0.365 |

NOTE: NST: Non-spatio-temporal; ST: Spatio-Temporal

## 4. Data analysis

In this section, for corn we compare the NST model and the ST model under the constraint that model planted acres must be larger than the FSA values. First, we look at the important hyper-parameters to show their importance in the models. Second, we look at the model estimates of the planted acres. In the summaries, we use posterior mean (PM), posterior standard deviation (PSD), posterior coefficient of variation (PCV) and 95% highest posterior density interval (HPDI) for the true state planted acres (*i.e.*, $\theta_i, i = 1, \ldots, \ell$). We consider only corn with $\ell = 48$ states. We also use maps and graphs to make more detailed comparisons.

### 4.1. Posterior inference of hyper-parameters

We look at posterior inference of some of the nuisance parameters. For example, the regression parameters contain important information; see Table 3.

Now, we discuss the results in Table 3. First, the Percent farmland irrigated has a negative effect on planted acres. Most of the speculative states for corn, except Nebraska, have little irrigation systems; California and the southern states have a lot of irrigation systems but less corn production. The value of cropland has a positive effect on planted acres, as it should. NCCPI has a positive effect on planted acres for corn. This must be true because better soil should lead to higher planted acres. This is also a good showing for the ST model, as under the NST model, while there is a large probability that $\beta_6$ is positive, the 95% HPDI contains 0. However, the Number of farms has a negative effect on planted acres. One possible explanation is the following. As the number of farms go up, one would expect smaller farms. In smaller farms, one would expect a larger variety of commodities, not fully dominated by corn.

We note that $\sigma^2$ is estimated very well under the ST model. It has a PCV of 2.21 under the NST model, but under the ST model, the PCV is 0.07, a huge improvement. The 95% HPDI for $\alpha_1$ is $(-23.68, 14.14)$ under the ST model, and it is good that $\alpha_1$ is not significant. Also, the 95% HPDI for $\alpha_2$ is $(.998, 1.009)$ under the ST model, and it is good that one is in it. (This is not true for the NST model.) This is important because it shows the power of the historical data. Here $\alpha_1$ and $\alpha_2$ are not identifiable in the models if there were no historical data. Another important point is that $\psi_1$ is closed to one in the ST model, but not so close under the NST model. Finally, the features of $p$, $z$ and $\rho$ are almost the same under both models. It is good that $\rho$ and $\psi_3$ are large under the ST model because it

**Table 3: Posterior summaries of hyper-parameters**

| | Non-spatial | | | | Spatial | | | |
|---|---|---|---|---|---|---|---|---|
| | *PM* | *PSD* | *PCV* | *HPDI* | *PM* | *PSD* | *PCV* | *HPDI* |
| $\beta_1$ | 2556.49 | 438.95 | 0.17 | (1750.30, 3336.09) | 2483.14 | 299.13 | 0.12 | (1868.99, 3044.76) |
| $\beta_2$ | -34.44 | 14.91 | -0.43 | (-60.13, -8.30) | -21.42 | 8.37 | -0.39 | (-37.19, -6.19) |
| $\beta_3$ | 0.56 | 0.49 | 0.89 | (-0.38, 1.41) | 0.01 | 0.36 | 19.63 | (-0.71, 0.72) |
| $\beta_4$ | 2771.19 | 207.03 | 0.07 | (2346.28, 3120.29) | 2932.2 | 105.1 | 0.04 | (2730.42, 3148.56) |
| $\beta_5$ | 3.24 | 3.48 | 1.07 | (-3.45, 9.31) | 6.27 | 2.12 | 0.34 | (2.18, 10.31) |
| $\beta_6$ | -0.03 | 0.004 | -0.15 | (-0.03, -0.02) | -0.03 | 0.002 | -0.08 | (-0.03, -0.02) |
| $\sigma^2$ | 58233 | 128555 | 2.21 | (16121, 167976) | 30922 | 2269 | 0.07 | (26396, 34931) |
| $\alpha_1$ | 3.55 | 1.19 | 0.34 | (2.01, 5.86) | -5.96 | 9.86 | -1.66 | (-23.68, 14.14) |
| $\alpha_2$ | 0.998 | 0 | 0 | (0.997, 0.998) | 1.002 | 0.004 | 0.004 | (0.998, 1.009) |
| $p$ | 0.38 | 0.09 | 0.24 | (0.19, 0.50) | 0.38 | 0.09 | 0.25 | (0.19, 0.50) |
| $z$ | 18.46 | 5.43 | 0.29 | (8.00, 28.00) | 18.38 | 5.54 | 0.3 | (6.00, 27.00) |
| $\rho$ | 0.96 | 0.003 | 0.003 | (0.95, 0.97) | 0.96 | 0.002 | 0.003 | (0.96, 0.97) |
| $\psi_1$ | 0.51 | 0.28 | 0.57 | (0.04, 0.98) | 0.99 | 0.01 | 0.01 | (0.97, 1.00) |
| $\psi_2$ | 0.51 | 0.28 | 0.57 | (0.02, 0.95) | 0.68 | 0.22 | 0.32 | (0.30, 1.00) |
| $\psi_3$ | – | – | – | (–,–) | 0.87 | 0.02 | 0.02 | (0.83, 0.89) |

NOTE: The five covariates are **Percent farmland irrigated**, Population density, **Value of cropland**, **National commodity crop production index (NCCPI)** and **Number of farms**. Here $z$ is the number of outliers. (The bolded covariates are important.)

shows that the CAR model has a significant effect.

### 4.2. Posterior inference for planted acres

In this section we compare the NST model and the ST model when we make posterior inference about planted acres under the constraint that the model planted acres are larger than the FSA planted acres.

In Table 4 we present posterior inference for the first thirteen states (in the order of state abbreviations), including small (*e.g.*, AZ, CT, FL) and some large (*e.g.*, IL, IN, IA) corn producing states. Apart from rounding, the constraints are satisfied in all states. The PMs are mostly similar and the PSDs under the spatial model are mostly smaller than those under the non-spatial model. This makes the PCVs under the spatial model mostly smaller than those under the non-spatial model, and therefore the 95% HPDIs are much shorter. These PCVs are smaller than the corresponding ones for the 'observed' data. Specifically, note that the gains in PCVs for CA, CO, FL and IL with unreliable data (larger CVp2). There are similar patterns for the other states, which are too numerous to list. We will look at all the states in greater detail using several plots (see below).

We now compare the spatial structure of the corn data under the constraint models. We have used the quintiles of the posterior means; note that the quintiles are not the same for the two sets of posterior means. In Figure 2, we show the map of the quintiles of planted acres. We can see some changes in these maps (CA, ID move from 2 to 3; AZ, NM move from 1 to 2; ND moves from 3 to 5; OH moves from 5 to 4, *etc.*). Otherwise, the two maps are mostly similar; however, the quintiles can hide the details, so we will discuss this further.

Figure 4 shows a plot of the posterior coefficients of variation (CVs) of the two models versus those of the observed data for the 48 states. We can see all the points are below the $45^o$ reference line, showing clearly that the two models provide improved reliability. We can also see most of the points corresponding to the non-spatial model are closer to the reference line than those from the spatial model, showing the estimates from the spatial model are more reliable. Those points, where a star and a dot are close together, correspond to the states with very large planted acres such as Iowa.

Figure 3 shows a plot of the posterior coefficients of variation (CVs) of the spatial model versus those of the non-spatial model for the 48 states. We can see all the points, except four (two very close), are below the $45^o$ reference line; the points falling on the reference line correspond to the states with large planted acres. This clearly shows that the spatial model provides improved reliability over the non-spatial model.

For completeness, we also look at the plot of PMs (Figure 5) and PSDs (Figure 6) for the spatial model versus the non-spatial model. For the PMs, it is really good that all of the points, except five of them, are nearly on the $45^o$ straight line through the origin. For the PSDs, it is also good that all of the points, except eight of them (five very close), are below the $45^o$ straight line through the origin. There is one of them in which the PSD is much lower under the ST model.

Integrating the (FSA, ASB) historical data into the models, which accommodate the survey data, appear to be important. Estimating the unknown FSA values in June of the current year is a reasonable thing to do. In general, the spatio-temporal (ST) model is better than the non-spatio-temporal (NST) model. The ST model fits the data better than the NST model. The constraint estimates from the ST model have smaller PCVs than those from the NST model.

These results show that the ST model provides higher precision and is more reliable than the NST model. Also the posterior means of the two models are very similar.

## 5. Concluding remarks

We have shown how to estimate planted acres for US states. This is on-going research and there are rapid changes under way as NASS pursued early estimates of planted acres, as early as June, and this is important for various reasons that we have discussed. As modernization and unification are under way at NASS, data integration is an important activity in this endeavor, and a lot of money and man power are put into it by NASS. Specifically, we have pointed out the struggles to find suitable statistical procedures in the initial stages. We have pointed out many challenges to get early estimates of planted acres and how to overcome some of them. Because of confidentiality, we have not used the real data in this paper, and the results presented may not be appropriate. As clearly described, the real data also have shortcomings, yet this project is extremely important to NASS.

In this paper, we have shown how to integrate a non-probability sample (FSA values) with a probability sample from a dual-frame survey (APS and JAS) to provide early estimates of planted acres for corn. One difficulty encountered is that the model estimates must be

**Table 4: Posterior summaries for planted acres (thousands) under the constraint**

| State | *fp1* | *p2* | *sp2* | *CVp2* | *PM* | *PSD* | *PCV* | *95% HPDI* |
|---|---|---|---|---|---|---|---|---|
| a. Non-spatial model | | | | | | | | |
| AL | 293 | 350 | 75 | 0.21 | 376.96 | 53.45 | 0.14 | (293.71, 478.25) |
| AZ | 88 | 95 | 19 | 0.2 | 106.25 | 12.45 | 0.12 | (87.98, 130.31) |
| AR | 733 | 750 | 103 | 0.14 | 818.61 | 62.06 | 0.08 | (733.64, 944.41) |
| CA | 401 | 470 | 161 | 0.34 | 570.15 | 113.02 | 0.2 | (401.17, 784.72) |
| CO | 1418 | 1400 | 646 | 0.46 | 1579.68 | 149.38 | 0.1 | (1418.42, 1879.06) |
| CT | 22 | 26 | 5 | 0.18 | 27.84 | 3.5 | 0.13 | (22.46, 34.41) |
| DE | 166 | 175 | 50 | 0.29 | 206.17 | 30.3 | 0.15 | (166.24, 264.16) |
| FL | 78 | 100 | 48 | 0.48 | 127.1 | 34.25 | 0.27 | (78.27, 194.12) |
| GA | 420 | 460 | 69 | 0.15 | 487.03 | 46.2 | 0.1 | (420.62, 578.79) |
| ID | 342 | 400 | 107 | 0.27 | 455.68 | 72.49 | 0.16 | (346.79, 585.72) |
| IL | 10465 | 11200 | 3511 | 0.31 | 10484.83 | 27.88 | 0 | (10464.58, 10542.13) |
| IN | 4988 | 5400 | 772 | 0.14 | 5028.64 | 51.56 | 0.01 | (4987.77, 5128.44) |
| IA | 12323 | 13100 | 3404 | 0.26 | 12338.54 | 19.79 | 0 | (12322.76, 12376.14) |
| b. Spatial Model | | | | | | | | |
| AL | 293 | 350 | 75 | 0.21 | 355.51 | 34.51 | 0.1 | (293.47, 417.46) |
| AZ | 88 | 95 | 19 | 0.2 | 99.2 | 7.63 | 0.08 | (87.99, 114.07) |
| AR | 733 | 750 | 103 | 0.14 | 792.53 | 41.53 | 0.05 | (733.44, 872.04) |
| CA | 401 | 470 | 161 | 0.34 | 551.69 | 72.56 | 0.13 | (405.20, 677.03) |
| CO | 1418 | 1400 | 646 | 0.46 | 1480.8 | 41.02 | 0.03 | (1420.05, 1538.34) |
| CT | 22 | 26 | 5 | 0.18 | 26.35 | 2.12 | 0.08 | (22.47, 30.22) |
| DE | 166 | 175 | 50 | 0.29 | 188.37 | 16.05 | 0.09 | (166.23, 217.93) |
| FL | 78 | 100 | 48 | 0.48 | 110.13 | 19.61 | 0.18 | (78.32, 144.76) |
| GA | 420 | 460 | 69 | 0.15 | 462.48 | 26.63 | 0.06 | (420.50, 511.14) |
| ID | 342 | 400 | 107 | 0.27 | 406.32 | 40.33 | 0.1 | (343.01, 480.69) |
| IL | 10465 | 11200 | 3511 | 0.31 | 10519.4 | 32.92 | 0 | (10464.97, 10572.29) |
| IN | 4988 | 5400 | 772 | 0.14 | 5066.48 | 45.83 | 0.01 | (4992.59, 5139.01) |
| IA | 12323 | 13100 | 3404 | 0.26 | 12368.6 | 27.1 | 0 | (12322.83, 12411.41) |

NOTE: fp1 is FSA planted acres, p2 is survey indications, sp2 is survey variance and CVp2 is survey coefficient of variation. The constraint specifies the model estimates must be larger than the FSA value.

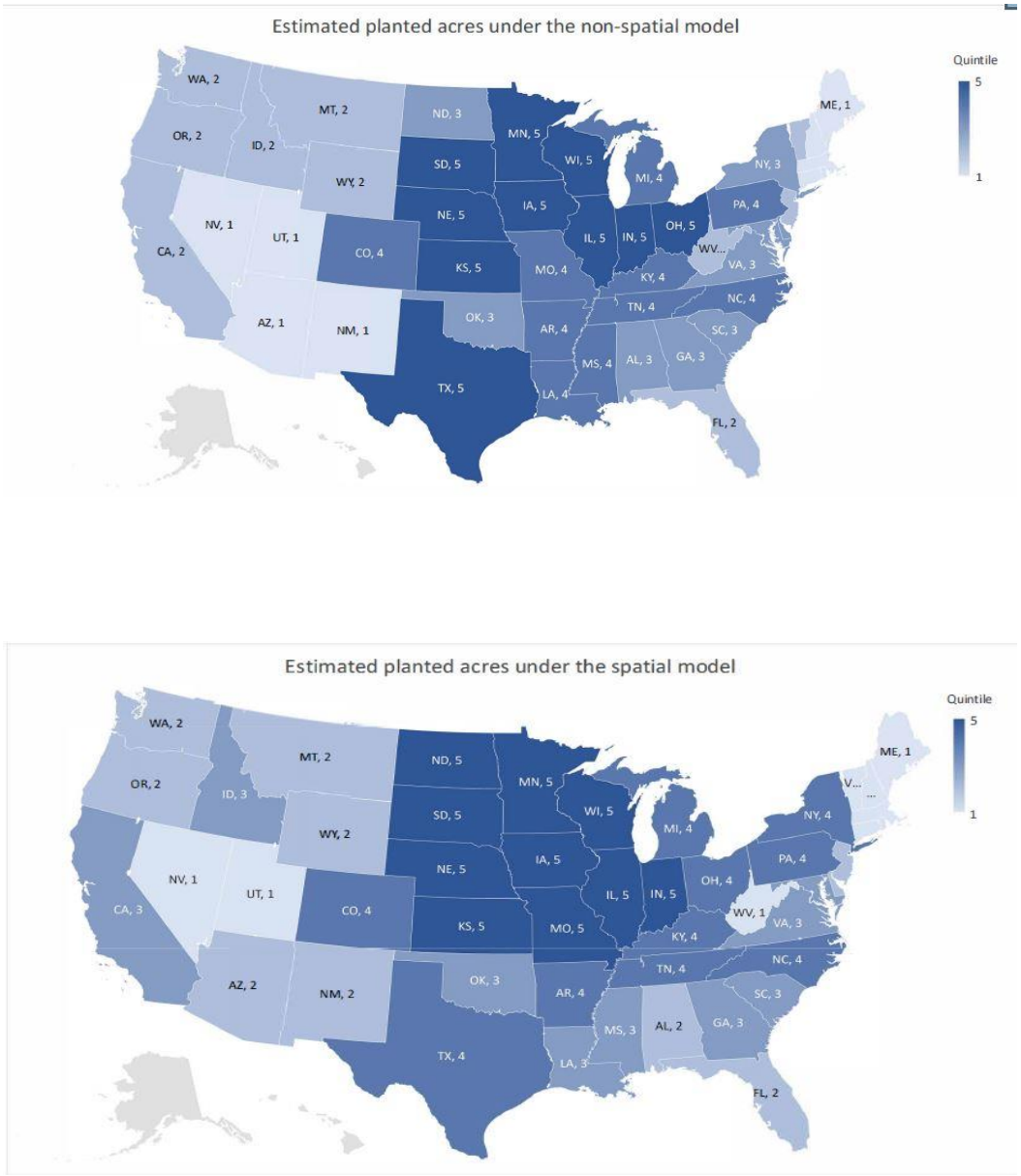


**Figure 2: Quintiles of posterior means of planted acres from the two models with constraint: The quintiles under the non-spatial (spatial) model are** 101 (92), 377 (356), 730 (668) **and** 3346 (3623), **and under the non-spatial (spatial) model, the minimum and maximum values are** 2.68 (2.32) **and** 12339 (12368).

larger than FSA values, which are unknown in June. We have provided two models for these data and we have demonstrated that the spatio-temporal model is a lot better than its non-spatio-temporal counterpart. While important indirect covariates, as used in this paper, are easily available, NASS has been pursuing more direct covariates such as temperature and precipitation, ethanol production capacity, and many others.

One would need to go down to lower level of disaggregation to accommodate variability. At the state level, there are actually a large number of records that go into the single number, thereby making variability relatively small because variance is generally inversely proportional to sample size. When survey indications are weighted up, there are no considerations of heterogeneity such as clustering (*e.g.*, counties) at intermediate levels, so that variability at the state level can be small. Young and Chen (2022) wrote, "Modeling at the state level is not always able to provide predictions of desired quality. Perhaps samples that provide valid estimates at lower geospatial scale should be considered; this will require major revisions in the current sample designs. Alternatively, if survey and non-survey data are linked at the farm level, then modeling could be conducted at that level."

It is now believed that modeling should be done at the level of Agricultural Statistics Districts (ASD); several ASDs might form a state. There are no ASD level survey indications and standard errors in June, so modeling is difficult to impossible at the ASD level; see Appendix A for a method to get ASD data from state data. Only state level indications and standard errors are available in June to NASS. We have been using the state level data to project backwards to the ASDs and the number counties within each ASD is used as the sample sizes (these are not presented) to get a rough idea of the indications and variances at the ASD level. A non-spatial model similar to the one discussed here is fit to the ASD level data, but now we need both an ASD level effect and a state level effect (so called sub-area or two-fold model). This will provide better state level model estimates. However, it is difficult to operationalize this model. At the ASD level, the NST model and the ST model are discussed in great detail in Nandram (2023), but again this second report is confidential. In addition, one may want to benchmark the states to the entire United States, but this is not attempted here. See Nandram, Ericulescu and Cruze (2019) for recent work on benchmarking.

A further problem of practical importance is the clustering of data at the state level, ASD level or county level. Many projects at NASS operates at county level such as cash rental rates and yield. The clustering does not have to be at geographical levels. For example, it does not have to be the case that the counties within a state have to form a cluster. Some counties in one state may be clustered with counties in another state. That is, there are unseen clusters among the sampling units (*e.g.*, counties), and these must be taken into consideration to avoid understating variability and biased estimates. Currently, this is on-going research activity in the Research and Development Division at NASS. Attempts are being made to accommodate this research activity for planted acres using the stick-breaking priors (Ishwaran and James, 2001); see Appendix B.

**Disclaimer, compliments, acknowledgements**

1. The findings and conclusions in this paper are those of the author and should not be

construed to represent any official USDA or US Government determination or policy.

2. The author thanks Neomi Guidin, Grace Yoon, Lu Chen, Ashley Lockwood, Valbona Bejeleri, Lindsay Drunasky, Andrew Dau, Chris Hawthorn and Linda Young for ongoing discussions at the National Agricultural Statistics Service (NASS).

3. The research methodology was supported by USDA's NASS.

4. Balgobin Nandram was supported by a grant from the Simons Foundation (#353953, Balgobin Nandram).

5. The paper was not written as part of Balgobin Nandram's normal duty at NASS, and a huge attempt was made to avoid confidential issues as much as possible.

6. Finally, I am grateful to the Organizing Committee, particularly Professor Vinod Gupta with his strong belief in me, for the invitation to present this Keynote Address. This was, indeed, a great honour for me.

## References

Berg, E., Im, J., Zhu, Z., Lewis-Beck, C., and Li, J. (2021). Integration of statistical and administrative agricultural data from Namibia. *Statistical Journal of the IAOS*, **37**, 557–578, DOI: 10.3233/SJI-200634.

Casella, G. and Berger, R. L. (2002). *Statistical Inference.* Second Edition. California: Duxbury, ISBN: 0-534-24312-6.

Chen, Y., Li, P., and Wu, C. (2020). Doubly robust inference with nonprobability survey samples. *Journal of the American Statistical Association*, **115**, 2011–2021, DOI: 10.1080/01621459.2019.1677241.

Chen, L., Nandram, B., and Cruze, N. (2022), Hierarchical Bayesian models with inequality constraints for US county estimates. *Journal of Official Statistics*, **38**, 709–732, DOI: 10.2478/jos-2020-0004.

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, DOI: 10.2307/2286322.

Fuller, W. A. (1987),. *Measurement Error Models.* New York: Wiley, ISBN: 0-471-86187-1.

Goyal, S., Datta, G. S., and Mandal, A. (2020). A hierarchical Bayes unit-level small area estimation model for normal mixture populations. *Sankhya B*, S1-S27, DOI: 10.1007/s13571-019-00216-8.

Ishwaran, H. and James, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, **96**, 161–172.

Janicki, R., Raim, M. A., Scott, H. H., and Maples, J. J. (2022). Bayesian nonparametric multivariate spatial mixture mixed effects models with application to American Community Survey special tabulations. *Annals of Applied Statistics*, **16**, 144–168, DOI: 10.1214/21-AOAS1494.

Kansas Farm Bureau (2020). *Counting crop acres NASS and WAOB vs. FSA*, 1–4.

Liu, J. S. (1994). The Collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem. *Journal of the American Statistical Association*, **89**, 958–966, DOI: 10.1080/01621459.1994.10476829.

Nandram, B. (2022). Temporal modeling of planted acres with spatial effects and covariates, *Technical Report*, Research Development Division, National Agricultural Statistics Service, USDA, 1–44.

Nandram, B. (2023). Temporal modeling of planted acres for Agricultural Statistics Districts with spatial effects and covariates, *Technical Report*, Research Development Division, National Agricultural Statistics Service, USDA, 1–77.

Nandram, B., Cruze, N. B., Erciulescu, A. L., and Chen, L. (2022). Bayesian small area models under inequality constraints with benchmarking and double shrinkage. *RDD Research Report, Number RDD-22-02, National Agricultural Statistics Service, USDA*, 1–41.

Nandram, B., Choi, J. W., and Liu, Y. (2021). Integration of non-probability and probability samples via survey weights. *International Journal of Statistics and Probability*, **10**, 5–21, 10.5539/ijsp.v10n6p5.

Nandram, B. and Rao, J. N. K. (2021). A Bayesian approach for integrating a small probability sample with a non-probability sample. *In JSM Proceedings, Survey Research Methods Section*, Alexandria, VA: American Statistical Association, 1568–1603.

Nandram, B. and Rao, J. N. K. (2023). Bayesian predictive inference when integrating a non-probability sample and a probability sample. *arXiv:2305.08997V1 [Stat.ME] 15 May 2023*, 1-35.

Nandram, B., Cruze, N. B., and Erciulescu, A. L. (2023). Bayesian small area models under inequality constraints with benchmarking and double shrinkage. *Survey Methodology* (in press), 1–30.

Nandram, B., Erciulescu, A. L., and Cruze, N. B. (2019). Bayesian benchmarking of the Fay-Herriot model using random deletion. *Survey Methodology*, **45**, 365–390.

National Academies of Sciences, Engineering, and Medicine (2017). *Improving Crop Estimates by Integrating Multiple Data Sources.* The National Academies Press: Washington, DC, USA, 2017.

Office of the Chief Economist (2019). *Update of 2019 FSA acreage data and FAQs on USDA acreage*, United States Department of Agriculture (USDA). Robert Johansson (Chief Economist) and Ashley Hungerford (Economist) for questions, August 27, 2019, pg. 0–7.

Rao, J. N. K. (2020). On making valid inferences by integrating data from surveys and other sources. *Sankhya B*, 3–33, DOI: 10.1007/s13571-020-00227-w.

Ridgeway, J. (2016). Computation of Gaussian orthant probabilities in high dimension. *Statistics and Computing*, **26**, 899–916, DOI: 10.1007/s11222-015-9578-1.

Ritter, C. and Tanner, M. A. (1992). Facilitating the Gibbs sampler: The Gibbs stopper and the griddy-Gibbs Sampler. *Journal of the American Statistical Association*, **87**, 861–868, DOI: 10.2307/2290225.

Tan, A. and Hobert, J. P. (2009). Block Gibbs sampling for Bayesian random effects models with improper priors: Convergence and regeneration. *Journal of Computational and Graphical Statistics*, **18**, 861–878, DOI: 10.1198/jcgs.2009.08153.

The National Agricultural Statistics Service (NASS) (2021 a). *Prospective Planting*, Agricultural Statistical Board, United States Department of Agriculture (USDA). Approved by Seth Meyer and Joseph L. Parson, March 31, 2021, ISSN: 1949-159X, pg. 1–36.

The National Agricultural Statistics Service (NASS) (2021 b). *Acreage*, Agricultural Sta-

tistical Board, United States Department of Agriculture (USDA). Approved by Seth Meyer and Joseph L. Parson, June 30, 2021, ISSN: 1949-1522, pg. 1–48.

Wang, J. C., Scott, H. H., Nandram, B., Barboza, W., Toto, C., and Anderson, E. (2012). A Bayesian approach to estimating agricultural yield based on multiple repeated surveys. *Journal of Agricultural, Biological, and Environmental Statistics*, **17**, 84–106, DOI: 10.107/513253-011-0067-5.

Young, L. J. and Chen, L. (2022). Using small area estimation to produce official statistics. *Stats*, **5**, 881–897, DOI: 10.3390/stats5030051.

## APPENDIX A

**How to get ASD level data from state level data?**

We consider a simple change of support (COS) analysis and we assume (normality is not required) that

$$\hat{\theta}_{ij} \overset{ind}{\sim} \text{Normal}(\theta_{ij}, \hat{\sigma}^2_{ij}), j = 1, \ldots, n_i, i = 1, \ldots, \ell,$$

where $n_i$ is the number of ASDs in the $i^{th}$ state (larger states have more ASDs). Let $m_{ij}$ denote the number of counties in the $j^{th}$ ASD. We do not know the $\hat{\theta}_{ij}$ and $\hat{\sigma}^2_{ij}$. However, note that $\sum_{j=1}^{n_i} \hat{\theta}_{ij} = \hat{\theta}_i$ and $\sum_{j=1}^{n_i} \hat{\sigma}^2_{ij} = \hat{\sigma}^2_i$ (assuming independence).

Specifically, we assume that $\hat{\theta}_{ij} \propto m_{ij}$, and this gives

$$\hat{\theta}_{ij} = \left\{ \frac{m_{ij}}{\sum_{j=1}^{n_i} m_{ij}} \right\} \hat{\theta}_i, j = 1, \ldots, n_i, i = 1, \ldots, \ell.$$

We also assume that $\hat{\sigma}^2_{ij} \propto {m_{ij}}^{-1}$, and this gives

$$\hat{\sigma}^2_{ij} = \left\{ \frac{m^{-1}_{ij}}{\sum_{j=1}^{n_i} m^{-1}_{ij}} \right\} \hat{\sigma}^2_i, j = 1, \ldots, n_i, i = 1, \ldots, \ell.$$

Both of these imputation procedures are reasonable because bigger states (i.e., planted acres of corn) will have larger $\hat{\theta}_i$ and smaller $\hat{\sigma}^2_i$.

Historical data, FSA values and ASB estimates, are available at county level. However, FSA values for the current year in June are not available and a similar procedure can be performed on the state values. Covariates can be used at the state level or jittered to get ASD level covariates. NASS will need to put in a large effort to get the covariates at the ASD level.

## APPENDIX B

### Basic stick-breaking distribution

For planted acres, the stick-breaking distribution for state estimates, $\hat{\theta}_i, i = 1, \ldots, \ell$, is

$$f(\hat{\theta}_i - \theta_i \mid \theta_i) = \sum_{s=1}^{\ell_o} p_s \text{Normal}(z_s, \hat{\sigma}_i^2), \ell_o \leq \ell,$$

where, given the $\theta_i$, the $\hat{\theta}_i - \theta_i$ are independent and identically distributed, the $p_s$ are stick-breaking weights, the $z_s$ are a random sample from a baseline distribution, and $\ell_o$ (unknown) is the number of clusters; see Ishwaran and James (2001). Therefore, it is true that

$$f(\hat{\theta}_i \mid \theta_i) = \sum_{s=1}^{\ell_o} p_s \text{Normal}(\theta_i + z_s, \hat{\sigma}_i^2), \ell_o \leq \ell,$$

and, given the $\theta_i$, the $\hat{\theta}_i$ are now independent, not identically distributed.

Introducing latent variables, this can be rewritten as

$$f(\hat{\theta}_i, d_i) = \prod_{s=1}^{\ell_o} [p_s \text{Normal}(\theta_i + z_s, \hat{\sigma}_i^2)]^{I(d_i = s)}, \ell_o \leq \ell,$$

where $d_i$ maps the $i^{th}$ state into a cluster and $I(d_i = s)$ is the indicator function.

Here the stick-breaking weights are

$$p_1 = \nu_1, p_s = \nu_s \prod_{r=1}^{s-1}(1 - \nu_r), s = 2, \ldots, \ell_o - 1, \ldots, p_{\ell_o} = \prod_{s=1}^{\ell_o - 1}(1 - \nu_s),$$

and for the two-parameter Pitman-Yor process, we use the prior,

$$\nu_s \overset{ind}{\sim} \text{Beta}\{1 - \delta_1, \frac{1 - \delta_2}{\delta_2} + (s - 1)\delta_1\}, s = 1, \ldots, \ell_o, 0 < \delta_1, \delta_2 < 1.$$

As for the $z_s$, we take

$$z_s \overset{ind}{\sim} \text{Normal}\{0, \frac{\rho}{1 - \rho}\sigma^2\}, s = 1, \ldots, \ell_o, 0 < \rho < 1.$$

It is also possible to assume a stick-breaking prior on the $\theta_i$.

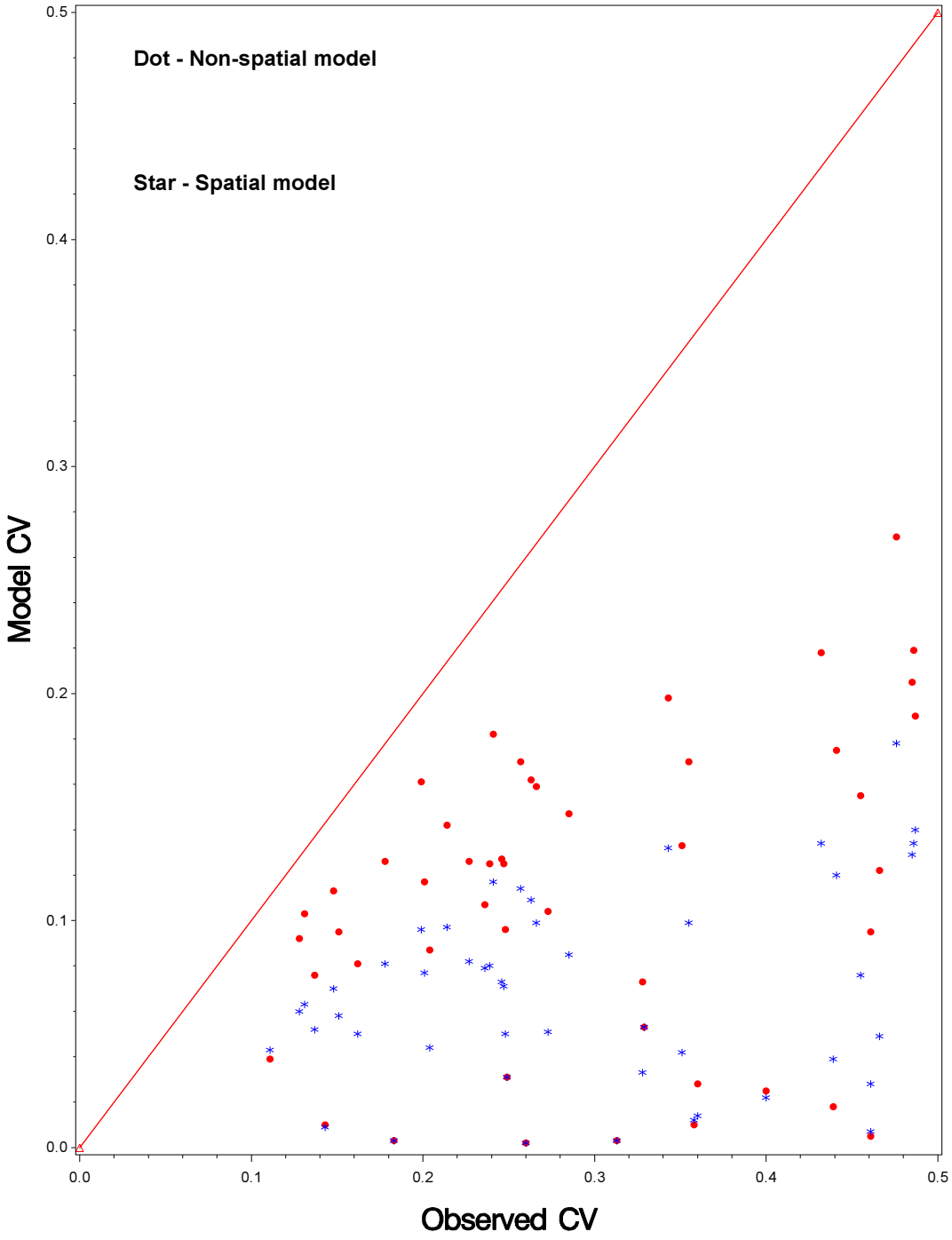


**Figure 3: Plots of the CVs of the two models versus the CVs of the observed data**

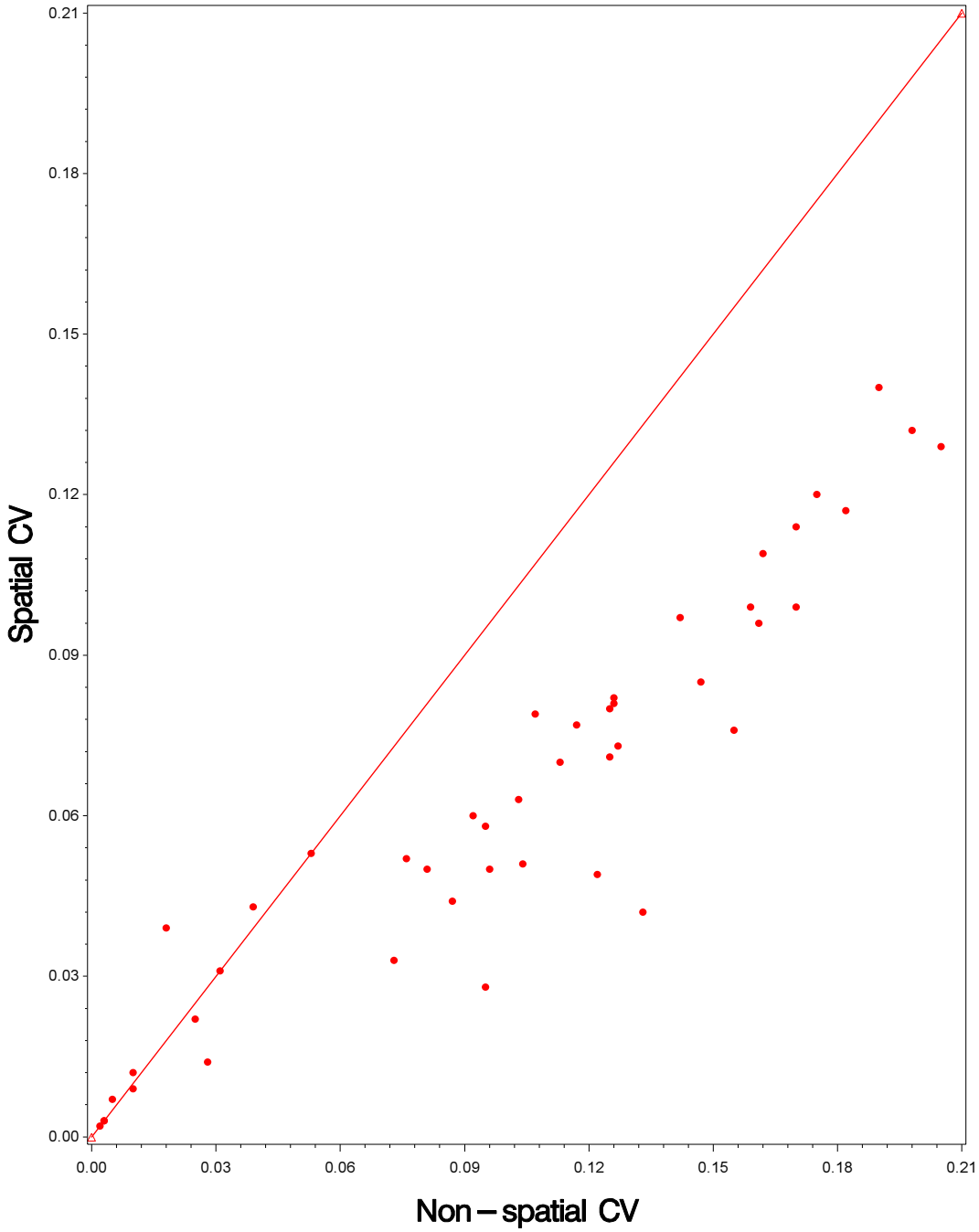


**Figure 4: Plots of the CVs of the spatial model versus those of the non-spatial model**

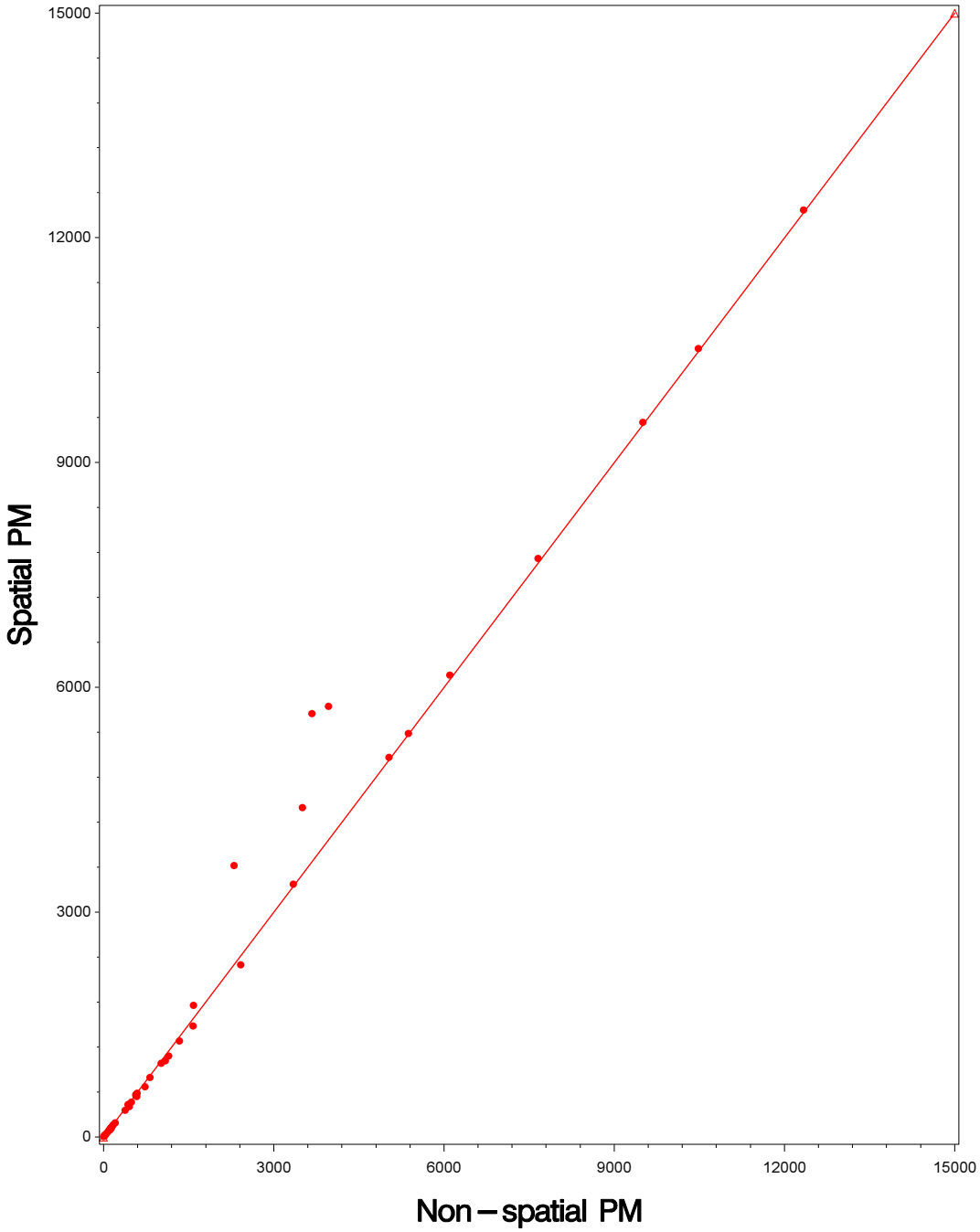


**Figure 5: Plots of the PMs of the spatial model versus those of the non-spatial model**

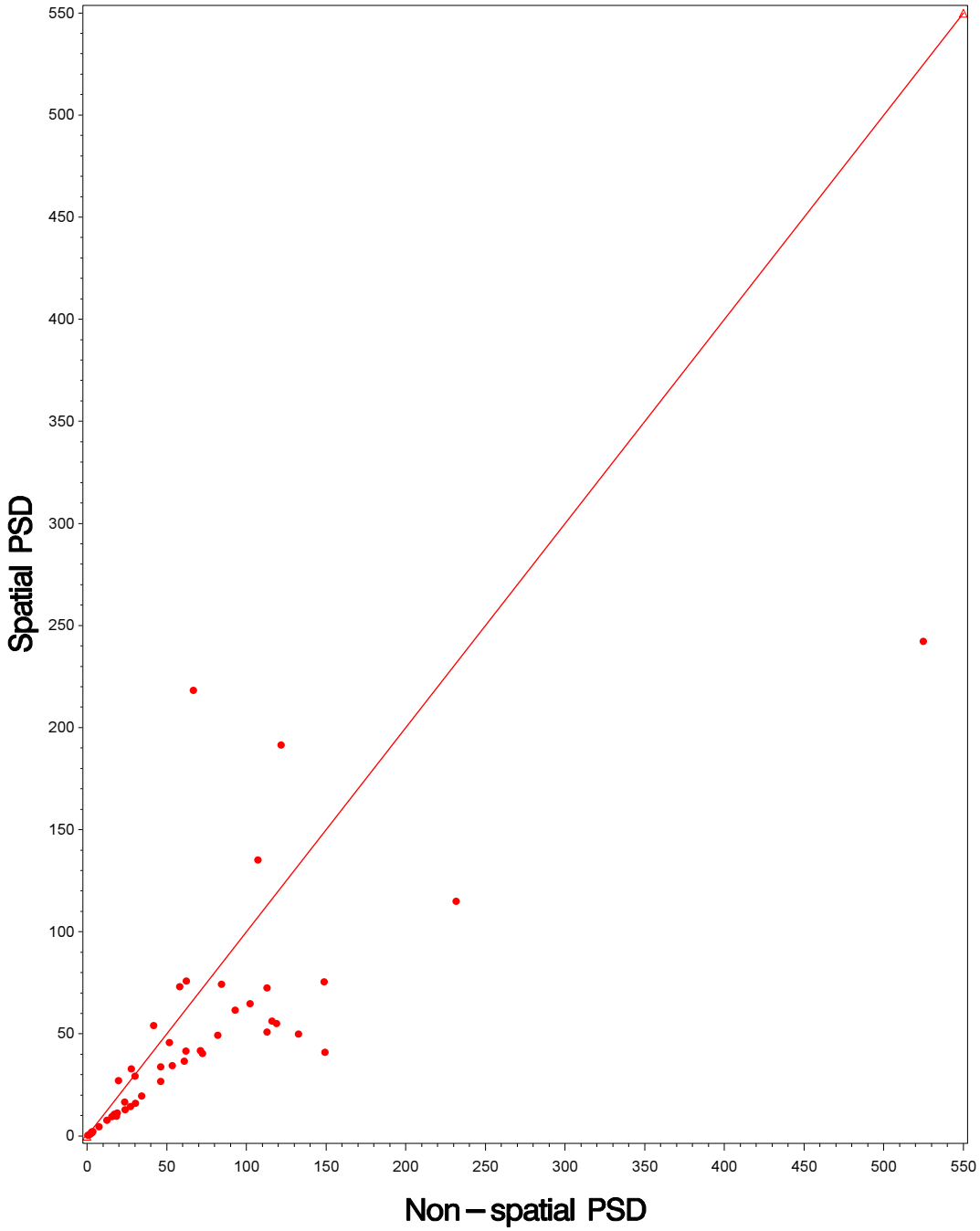


**Figure 6: Plots of the PSDs of the spatial model versus those of the non-spatial model**

## APPENDIX C

### A list of useful abbreviations

| Abbreviations | Meanings |
|---|---|
| USDA | United States Department of Agriculture |
| NASS | National Agricultural Statistics Service |
| FSA | Farm Service Agency |
| RDD | Research Development Division |
| ASB | Agricultural Statistics Board |
| APS | Agricultural Production Survey |
| JAS | June Area Survey |
| ASD | Agricultural Statistics District |

NOTE: NASS and FSA are two of the agencies of USDA, and RDD is a division of NASS. APS and JAS are the two surveys. All estimates are approved by the ASB before publication.