A model-based approach to forecasting corn and soybean yields

Daniel W. Adrian*

Abstract

The National Agricultural Statistics Service (NASS) publishes forecasts and estimates of yields for several major crops. The yield forecasts have important economic implications through their effects on market prices and the World Agricultural Outlook Board's supply and demand estimates. NASS collects information for forecasting corn yields from two sets of monthly surveys. A large scale farmer interview survey is conducted after the corn is harvested to determine the final yield estimate. Historically, yield forecasts are established by the Agricultural Statistics Board, a group of commodity specialists that convenes monthly and synthesizes the survey results with external sources information. Because of the subjective nature of the Board process, the traditional estimation procedures are not reproducible, are difficult to document precisely, and do not provide a measure of uncertainty. A model-based approach to corn and soybean yield forecasting is proposed as an alternative to the Board process. The model incorporates the direct survey estimates as well as external sources of information and produces a measure of statistical efficiency.

Key Words: Bayesian hierarchical model; Composite estimation; Model-based estimation; Survey sampling.

1. Introduction

Each month, the U.S. Department of Agriculture (USDA) publishes two reports which heavily influence crop prices by "defining the fundamental conditions in commodity markets" (Vogel and Bange 1999, p. 3). The first, the Crop Production Report, which is prepared by the National Agricultural Statistics Service (NASS), an agency within the USDA, contains estimates of harvested and planted area, yield, and production for a multitude of U.S. crops at the national and state levels. This crop production information is utilized in preparing the second report, the World Agricultural Supply and Demand Estimates (WASDE), which forecasts crop supply and utilization in the U.S. and worldwide. Due to the market-sensitive nature of these reports, they are prepared under "lockup": those involved are literally locked in an area without internet and telephone access from midnight until the report is released at 8:30 a.m. Eastern time. We focus on the Crop Production Report – specifically, the production of corn and soybeans, the two largest U.S. crops. Production estimates of these two crops are published in tandem: "in-season" estimates are published for the August through November reports, and the end-of-season estimates are determined in January.

Production refers to the total harvested "fruit," measured (domestically) in bushels, which, for corn and soybeans, refers to kernels and beans after removal from ear and pod, respectively. The bushel, though traditionally a unit of (dry) volume, refers to standard weights; bushels of corn and soybeans at standard moisture levels weigh 56 and 60 pounds, respectively. NASS computes production as harvested area multiplied by yield, where area is measured in acres and yield is measured in bushels per acre. An acre is roughly the size of an American football field, and 640 acres equal one square mile.

Of yield and area, it is widely recognized that yield is more difficult to measure; thus, yield estimation is the main component of the determination of production. While there is

*National Agricultural Statistics Service, Research & Development Division, Suite 305, 3251 Old Lee Highway, Fairfax, VA 22030

a large amount of area to estimate – corn and soybeans contribute roughly 84 million and 74 million harvested acres, respectively (USDA NASS 2012a), the estimation is supported by NASS's largest survey, the June Agricultural Survey (JAS); the JAS collects the planted area of over 125,000 farmers per year. Further, the harvested area estimates are highly correlated to planted acres, with the former being roughly 98% of the latter. On the other hand, the true yield is not known until harvest, and yield is difficult to predict in the early stages of the season. In the following, we review NASS's yield estimation methods.

Crop yield estimation at NASS is primarily based upon three probability-based surveys: the Agricultural Yield Survey (AYS), Objective Yield Survey (OYS), and December Agricultural Survey (DAS). The AYS and OYS, both of which are panel surveys subsampled from the JAS, are conducted monthly for the in-season reports from August through November, and the OYS also occurs in December. The AYS is a farmer interview survey, conducted in nearly every state, in which producers are asked for their expected crop yields. The Objective Yield Survey, so-named due its independence from farmer response, is a field measurement survey in which enumerators take commodity-specific measurements (*e.g.* number of ears of corn) from randomly selected plots within the selected fields. Due to its restrictive expense, the OYS is only undertaken in the highest-producing states; NASS refers to these groups of states as speculative (or "spec") regions due to their market influence. The DAS is a farmer interview survey like the AYS, but it is only conducted in December after the harvest is complete and has a much larger sample size than either the AYS or OYS. As a result, of the three surveys, the DAS provides the most reliable indication of the end-of-season yield. After weighting the survey records to adjust for selection probabilities and response, yield indications and standard errors are produced for each survey. At NASS, these direct survey estimates are referred to as "indications" instead of estimates as the latter term is reserved for published numbers.

Indications from the OYS and AYS improve as estimators of the end-of-season yield as the growing season progresses. For instance, early-season AYS responses may be well-described as opinion due to the lack of information upon which they are based, but, later on, AYS responses are based upon actual harvest numbers. The same is true for the OYS. Consider the OY model for corn yield, which considers the grain weight per ear (USDA NASS 2006). In the early season when the kernels are not fully formed, OY models predict the grain weight through historical averages; however, when the corn is mature later in the season, the grain weight is measured directly from harvested ears.

It is widely recognized at NASS that the OYS and AYS indications are biased as estimators of the end-of-season yield; specifically, the OYS indications have a positive bias while the AYS indications have a negative bias. As a result, these indications are adjusted for bias, where the biases are estimated from historical differences between the indications and end-of-season yield estimates. The source of these biases is not completely understood. The negative AY bias is generally attributed to the tendency for farmers to be pessimistic and conservative in outlook. The positive OY bias is attributed to the OY model specification (Warren 1985), but the details are not well understood. In contrast, the DAS indications are treated as unbiased due to their collection after harvest completion.

In addition to the survey indications, NASS considers various sources of auxiliary information. For example, NASS monitors weather conditions such as temperature and precipitation. Specifically, high corn yields are commonly associated with below-normal temperatures and above-normal precipitation in July and August (Thompson 1986) while soybean yields are particularly sensitive to August precipitation (Tannura, Irwin, and Good 2008). Further, NASS's weekly Crop Progress and Conditions survey provides crop condition ratings in terms of five categories from very poor to excellent. NASS also utilizes satellite information such as the Normalized Difference Vegetative Index (NDVI), a mea-

sure of biomass density which has been shown to be related to yield (Doraiswamy et al 2005). Last, a linear time trend is commonly used to model the increasing yields brought about by technological innovations; this effect has been most pronounced for corn yields, which have ten-year averages of 46, 72, 92, 109, 125, and 149 bushels per acre beginning in 1950 (USDA NASS, 2012b).

The final published estimates are determined by the Agricultural Statistics Board (ASB), a group of commodity experts which convenes during lockup. The ASB sets yield estimates for states comprising the spec region during lockup; the non-spec state yields are determined before the meeting. Recall that the spec regions are the groups of states employing the OYS; the regions for corn and soybeans include 10 and 11 states, respectively, and account for 84 and 83 percent of U.S. production (USDA NASS 2012a). In determining the yield estimates for spec region states, the ASB follows a top-down approach: for each commodity, it first sets yields for the speculative region as a whole and then sets yields for its component states subject to the restriction that the harvested-acres-weighted average of the state-level yield estimates equals the regional yield estimate. For the first task, the ASB considers spec-region-level indications, which are aggregated from the state level as a harvested-acres-weighted average. This top-down approach is justified by the belief that estimates are strongest at the regional level where the sample size is largest. Note, however, that this approach does not specify an explicit estimation method.

For this reason, the ASB has been criticized for being subjective. Other than calculating the published yield as a simple average of each Board member's estimate, the ASB does not utilize explicit mathematical rules to synthesize the variety of yield information sources. As previously mentioned, early-season yield estimation is a difficult task, and this lack of transparency grants them flexibility to adapt to changing and uncertain crop conditions. However, as a result, the emphases placed on different features of the data will inevitably vary from person to person and depend upon the composition of the Board (See Caristi (2012), who studies the dispersion of the Board members' estimates over time.) Thus, reproducibility is an issue: if the ASB is provided with the same information on two different occasions, there is no guarantee that it would arrive at the same yield estimate. Further, no statistical measure of uncertainty such as a standard error is provided. Although a reliability measure is published for the forecasted estimates, it is only based on historical differences between the projected estimates and the end-of-season estimates. These criticisms are motivated by OMB Standard 4.1, which says that "Agencies must use accepted theory and methods when deriving ... model-based estimates and projections that use survey data" and that "error estimates must be calculated" (OMB 2006, p. 20).

NASS has made efforts to address these criticisms. Keller and Olkin (2002) developed a composite yield estimator which computes a weighted average of bias-adjusted AY and OY indications. Although this methodology removes subjectivity from the yield estimation process and provides a standard error, it does not incorporate the sources of auxiliary information available to the ASB.

To this end, in a cooperative agreement with the National Institute of Statistical Sciences (NISS), NISS-NASS researchers developed a Bayesian hierarchical model for the spec-region-level corn yield (Wang et al 2011). Because this model provides the foundations for this paper, we review this model in detail, describing the modeling levels starting at the top. The top level models the finite sampling processes which produce the survey indications; that is, the survey indications differ from their superpopulation means due to sampling errors, which are quantified according to the given standard errors. Below this level, the superpopulation means of the OYS and AYS indications are linked with the true end-of-season yields through bias parameters and forecasting errors. The DAS indications, however, are assumed unbiased for the true yields with no forecasting error. At both of the

previous levels, the model recognizes correlations between monthly OYS and AYS indications within the same year and survey due to the repeated panel designs. At the lowest level, the underlying true yields follow a linear regression model in terms of several auxiliary covariates. Yield estimates are computed via MCMC sampling from the posterior distribution defined by this model, for which a Gibbs sampling algorithm is described.

My NASS colleagues and I have modified and extended this hierarchical Bayesian modeling approach. First, we pared down this model for spec-region-level yield to a more parsimonious version. Our model retains most of the previous features, except that now only the current month OYS and AYS indications are utilized. Further, we develop a model for the individual spec region states – herein referred to as the state-level model – that applies the ASB's top-down approach. That is, after obtaining the spec-region-level yield estimate, we apply to the method of Nandram and Sayit (2011) so that the harvested-acres-weighted average of the state-level model yield estimates equals the regional estimate.

The article proceeds as follows. Section 2 provides the necessary background information about the survey indications and auxiliary covariates. Section 3 introduces the methodology behind the spec-region- and state-level yield models. We provide concluding remarks in Section 4.

2. Survey indications and auxiliary covariates

In this section, we provide the necessary background information about the data. We begin by discussing the temporal and geographical supports of the survey indications. Regarding the former, the indications cover different months and years depending on survey. As previously mentioned, the OYS is employed monthly from August through December, the AYS monthly from August through November, and the DAS once in December. In addition, indications for the OYS, DAS, and AYS are not available before 1993, 1996, and 2001, respectively: the first two are due to changes in NASS processing environments and the latter is due to a change in sampling methodology. Further, recall that while the AYS and DAS are taken in almost all states, the OYS is only employed in spec region states. There are currently 10 and 11 states in the corn and soybean spec regions, respectively, and nine states are in both: Illinois, Indiana, Kansas, Minnesota, Missouri, Nebraska, Ohio, and South Dakota. The addition of Wisconsin completes the current corn spec region, while Arkansas and North Dakota round out the soybean region. Three states were added to each spec region in 2004, which included Kansas and South Dakota for both commodities, Missouri for corn, and North Dakota for soybeans. Arkansas was also added to the soybean spec region in 2004, but, unlike the other added states, it was part of the region prior to and including 2001.

Next, we introduce the utilized set of auxiliary covariates. We use monthly, state-level temperature and precipitation data provided by NOAA (NOAA NESDIS 2012), which we denote by `TMPmm` and `PCPmm`, respectively, with `mm` referring to the month number. We use July data in the corn yield model and both July and August data for soybeans. We also use the state-level crop condition ratings (in percent) given by the weekly Crop Progress and Condition Survey. Specifically, we consider the sum of the percentages for the "good" and "excellent" ratings, which we denote by `AGRww`, where `ww` refers to the week number; we use the 30th and 34th week ratings for corn and soybeans, respectively, which occur near the last week of July and August. The covariates are available for all years.

2.1 Harvested-acres-weighted aggregation of state-level yields

As previously mentioned, NASS aggregates state-level yield indications and estimates to the spec region level according to a harvested-acres-weighted average. After introducing some notation, we show that this practice follows directly from the definition of production. Denote the production, yield, and harvested acreage in year t and state l by p_{tl}, μ_{tl}, and h_{tl}, respectively, and the corresponding spec-region-level quantities by p_t, μ_t, and h_t. Further, denote the year-t proportion of spec-region-level harvested acreage coming from state l by $w_{tl} \equiv h_{tl}/h_t$. Now, $p_t = \sum_{l=1}^{L} p_{tl} = \sum_{l=1}^{L} h_{tl}\mu_{tl}$ by the definition of production. Dividing both sides by h_t gives

$$\mu_t = \sum_{l=1}^{L} w_{tl}\mu_{tl}. \tag{1}$$

Though not mathematically complex, this relationship motivates the methodology for the state-level yield model.

2.2 Argument for using only current month indications

As mentioned in the introduction, the Wang et al (2011) model uses all monthly OYS and AYS indications. For example, in calculating the October Crop Production Report forecast, this model would utilize the August, September, and October OYS and AYS indications. We argue that only the OYS/AYS indications from the current month (per our example: October only) should be used. As previously mentioned, the record-level AYS and OYS data becomes more predictive of the end-of-season yield as the season progresses: while early season data are based upon farmer opinion and historical averages, late season indications use current-year harvest data. Furthermore, the OYS and AYS are panel surveys, meaning the records represent the same farmers and fields in each month of the survey. Thus, we argue that it does not make sense to include outdated indications in the model when more current information is available.

3. Methodology

3.1 Notation

Due to the general nature of the model, our notation is not commodity-specific. Denote the survey indication from survey k, year t, month m, and state l by y_{ktml} and its (given) sampling standard error by s_{ktml}, where $k = O, A, D$ for OYS, AYS, and DAS, $t \in \{1, 2, \ldots, T\}$, T representing the forecasting year, $m \in \{8, 9, 10, 11, 12\}$ for August through December, and $l \in \{1, \ldots, L\}$. To be more specific, the index ranges are $t = t_k, t_k + 1, \ldots, T$, where t_k represents the first year of available data for survey k; $m = 8, 9, \ldots, 12$ for OYS, $m = 8, 9, 10, 11$ for AYS, and $m = 12$ for DAS; and $l = 1, \ldots, L(t)$ due to the changing definition of the spec region. Further, we denote the vector of covariates for year t and state l by $\boldsymbol{z}_{tl}$. The corresponding spec-region-level quantities are aggregated from the state-level quantities according harvested-acres-weighted averages: the spec-region-level indications, standard errors, and covariates are computed as $y_{ktm} = \sum_{l=1}^{L(t)} w_{tl} y_{ktml}$, $s^2_{ktm} = \sum_{l=1}^{L(t)} w^2_{tl} s^2_{ktml}$, and $\boldsymbol{z}_t = \sum_{l=1}^{L(t)} w_{tl}\boldsymbol{z}_{tl}$, where the variance aggregation assumes independence between the states.

Continuing, we let μ_t and μ_{tl} define the spec-region- and state-level end-of-season yields and denote the bias and non-sampling variance associated with the spec-region-level indication for survey k and month m by b_{km} and σ^2_{km}; the corresponding state-l quanti-

ties are b_{kml} σ^2_{kml}. Last, we denote the forecast month by m^*, $m^* = 8, \ldots, 12$, which correspond to the August through November yield estimates and the end-of-year estimate.

3.2 Spec-region-level model

For forecasting month m^*, the spec-region-level yield model is given by

$$\begin{aligned} y_{ktm^*}|\mu_t &\sim \text{indep } N(\mu_t + b_{km^*}, s^2_{ktm^*} + \sigma^2_{km^*}),\ t_k \le t \le T,\ k = O, A, && (2)\\ y_{Dt,12}|\mu_t &\sim \text{indep } N(\mu_t, s^2_{Dt,12}),\ t_D \le t \le T(m^*), && (3)\\ \mu_t &\sim \text{indep } N(\boldsymbol{z}_t'\boldsymbol{\beta}, \sigma^2_\eta), 1 \le t \le T, && (4) \end{aligned}$$

where $T(m^*) = T$ for $m^* = 12$ and $T - 1$ otherwise, and $\boldsymbol{z}_t$ depends on m^* and the commodity. For the corn yield model, $\boldsymbol{z}_t = (1, t, \texttt{PCP07}_t, \texttt{Temp07}_t, \texttt{AGR30}_t)'$ for all m^*; for soybeans, the same variables are used for $m^* = 8$, but $\boldsymbol{z}_t = (1, t, \texttt{PCP08}_t, \texttt{Temp08}_t, \texttt{AGR34}_t)'$ for $m^* > 8$ because the August information is more predictive of the soybean yield. Note that (2)-(4) follows the Wang et al (2011) model except for the following: first, as previously mentioned, we only consider OYS and AYS indications at month m^*. Further, we integrate out the superpopulation mean parameters and use the condition rating covariate instead of planting progress. The application of diffuse prior distributions, which is described in Appendix A.1, completes the model specification.

We employ MCMC simulation via a Gibbs sampling algorithm (Gelman et al 2004) for parameter estimation. We show only the full conditional distribution of μ_T, the parameter of interest, and leave the remaining parameters to Appendix A.2. Denoting the set of other parameters by $\boldsymbol{\Omega}$,

$$\mu_T|\boldsymbol{\Omega} \sim N(\Delta_2/\Delta_1, 1/\Delta_1), \tag{5}$$

$$\begin{aligned} \Delta_1 &= \sum_{k=O,A} \frac{1}{\sigma^2_{km^*} + s^2_{kTm^*}} + \frac{I_{\{m^*=12\}}}{s^2_{DT,12}} + \frac{1}{\sigma^2_\eta},\\ \Delta_2 &= \sum_{k=O,A} \frac{y_{kTm^*} - b_{km^*}}{\sigma^2_{km^*} + s^2_{kTm^*}} + \frac{I_{\{m^*=12\}} y_{DT,12}}{s^2_{DT,12}} + \frac{\boldsymbol{z}_T'\boldsymbol{\beta}}{\sigma^2_\eta}, \end{aligned}$$

where $I_{\{.\}}$ is the indicator function with I_A equal to 1 if A is true and 0 otherwise. The yield estimate $\hat{\mu}_T$ is the posterior mean, the mean of the MCMC sample for μ_T after the "burn-in" period has been removed. Similarly, the standard error of $\hat{\mu}_T$ is the standard deviation of this sample.

3.3 State-level model

The state-level model follows the same structure as the regional-level model. For month m^*, it is given by

$$\begin{aligned} y_{ktm^*l}|\mu_{tl} &\sim \text{indep } N(\mu_{tl} + b_{km^*l}, s^2_{ktm^*l} + \sigma^2_{km^*l}),\ k = O, A && (6)\\ y_{Dt,12,l}|\mu_t &\sim \text{indep } N(\mu_{tl}, s^2_{Dt,12,l}), && (7)\\ \mu_{tl} &\sim \text{indep } N(\boldsymbol{z}_{tl}'\boldsymbol{\beta}_l, \sigma^2_{\eta l}) && (8) \end{aligned}$$

over the ranges of t and l. The parameters have diffuse priors such as those given in Appendix A.1. The state-level model yield estimates $\hat{\mu}_{tl}$, like the spec-region-level model estimates $\hat{\mu}_t$, are the means of MCMC samples computed via a Gibbs sampling algorithm. Although (1) holds for the parameters μ_t and μ_{tl}, the same relationship does not necessarily

hold for the estimates $\hat{\mu}_t$ and $\hat{\mu}_{tl}$. As a result, we refer to (6)-(8) as the "unconstrained" model.

Because the incompatibility of the state- and regional-level yield estimates is not acceptable, we enforce this relation by specifying a constrained state-level model. In an adaptation of the class of models developed by Nandram and Sayit (2011), we augment (6)-(8) with the constraint (1) for each t. In doing so, we use the posterior distribution of μ_t from the spec-region-level model as its prior distribution for the constrained state-level model. Thus, the constrained state-level model may be thought of as a hierarchical model with regional and state modeling levels – given by (2)-(4) and (6)-(8), respectively – linked by (1). In practice, this involves performing MCMC sampling for the spec-region-level model first and then inputting these μ_t iterates into the Gibbs sampling algorithm for the state-level model.

For each t, the full conditional distribution of $\boldsymbol{\mu}_{t\cdot} \equiv (\mu_{t1}, \ldots, \mu_{t,L(t)})$ under the constrained state model is obtained by conditioning its (full conditional) distribution under the unconstrained model on (1). The unconstrained-model distribution is given by

$$\begin{aligned}
\boldsymbol{\mu}_{t\cdot}|\boldsymbol{\Omega} &\sim \text{indep } MVN(\underset{1\le l\le L(t)}{\text{vec}}\left(\frac{\Delta_{2tl}}{\Delta_{1tl}}\right), \underset{1\le l\le L(t)}{\text{diag}}\left(\frac{1}{\Delta_{1tl}}\right)), \qquad (9)\\
\Delta_{1tl} &= \sum_{k=O,A} \frac{I_{\{t>t_k\}}}{\sigma^2_{km^*l} + s^2_{ktm^*l}} + \frac{I_{\{m^*=12\}}}{s^2_{Dt,12,l}} + \frac{1}{\sigma^2_{\eta l}},\\
\Delta_{2tl} &= \sum_{k=O,A} \frac{y_{ktm^*l} - b_{km^*l}}{\sigma^2_{km^*l} + s^2_{ktm^*l}} + \frac{I_{\{m^*=12\}} y_{Dt,12,l}}{s^2_{Dt,12,l}} + \frac{\boldsymbol{z}'_{tl}\boldsymbol{\beta}_l}{\sigma^2_{\eta l}},
\end{aligned}$$

where $\text{vec}(\cdot)$ and $\text{diag}(\cdot)$ represent vectors and diagonal matrices (with entries indexed by l). Then, by properties of the multivariate normal distribution, it can be shown that conditioning (9) on (1) gives the distribution defined by $(\mu_{t1}, \ldots, \mu_{t,L(t)-1}) \sim N(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\Sigma}})$ and $\mu_{t,L(t)} = (\mu_t - \sum_{l=1}^{L(t)-1} w_{tl}\mu_{tl})/w_{t,L(t)}$, where

$$\begin{aligned}
\bar{\boldsymbol{\mu}} &= \underset{1\le l\le L(t)-1}{\text{vec}}\left(\frac{\Delta_{2tl} - a w_{tl}}{\Delta_{1tl}}\right), \qquad (10)\\
\bar{\boldsymbol{\Sigma}} &= \underset{1\le l\le L(t)-1}{\text{diag}}\left(\frac{1}{\Delta_{1tl}}\right) - b \underset{1\le l,l'\le L(t)-1}{\text{mat}}\left(\frac{w_{tl}w_{tl'}}{\Delta_{1tl}\Delta_{1tl'}}\right), \qquad (11)
\end{aligned}$$

where $\text{mat}(\cdot)$ represents a matrix (with rows and columns indexed by l and l', respectively), and the scalars a and b are given by $b = [\sum_{l=1}^{L(t)} w_{tl}^2/\Delta_{1tl}]^{-1}$ and $a = b[\sum_{l=1}^{L(t)} w_{tl}\Delta_{2tl}/\Delta_{1tl} - \mu_t]$. For sake of brevity, we omit the full conditional distributions of the other state-level model parameters, which are similar to those in Appendix A.2.

3.4 Weighted average decompositions of yield estimates

In this section, we present useful interpretations of the spec-region- and state-level yield estimates as weighted averages of estimates coming from each survey indication and the auxiliary covariates. For instance, the year-T and month m^* spec-region-level yield estimate may be approximated as

$$\hat{\mu}_T \approx \sum_{i=O,A,D,C} c_i \tilde{\mu}_i, \qquad (12)$$

where $\tilde{\mu}_i$, for $i = O, A, D, C$, are the separate estimate components coming from the OYS, AYS, DAS, and auxiliary covariates, respectively, and c_i are their weights such that

$\sum_i c_i = 1$. The approximation (12) is computed from the mean of the full conditional distribution (5), applying the posterior means of the parameters $\boldsymbol{\Omega}$. (The approximation falls within one-tenth bushel of $\hat{\mu}_t$.) The estimate components for the OYS, AYS, and DAS are the bias-adjusted indications $\tilde{\mu}_i = y_{iTm^*} - \hat{b}_{im^*}$, $i = O, A$ ($\hat{b}_{im^*}$ denoting the posterior mean of b_{im^*}), and $\tilde{\mu}_D = y_{DT,12}$; the estimate component from the auxiliary covariates is the fitted regression line $\tilde{\mu}_C = \boldsymbol{z}_T'\hat{\boldsymbol{\beta}}$. Further, the weights c_i are proportional to the inverse variances for each data source; specifically, $c_i = \kappa_i / \sum_i \kappa_i$, where $\kappa_i = 1/(\hat{\sigma}^2_{im^*} + s^2_{iTm^*})$ for $i = O, A$, $\kappa_C = 1/\hat{\sigma}^2_\eta$, and $\kappa_D = I_{\{m^*=12\}}/s^2_{DT,12}$.

A similar interpretation is available for the state-level yield estimates. The year-T and month m^* state-level yield estimates, for $l = 1, \ldots, L$, may be approximated as

$$\hat{\mu}_{Tl} \approx \sum_{i=O,A,D,C} c_{il}\tilde{\mu}_{il} + d_l, \tag{13}$$

where $\tilde{\mu}_{il}$ and c_{il} are the state-l estimate components and weights such that $\sum_i c_{il} = 1$ for each l. Further, d_l represents the adjustment to the state-l unconstrained-model yield estimate in order that the state-level yield estimates are compatible with the regional-model yield estimate. Using (9) and the same approximation method as before, the estimate components are $\tilde{\mu}_{il} = y_{iTm^*l} - \hat{b}_{im^*l}$, $i = O, A$, $\tilde{\mu}_{Dl} = y_{DTm^*l}$, and $\tilde{\mu}_{Cl} = \boldsymbol{z}_{tl}'\hat{\boldsymbol{\beta}}_l$; the weights are given by $c_{il} = \kappa_{il}/\sum_i \kappa_{il}$, where $\kappa_{il} = 1/(\hat{\sigma}^2_{im^*l} + s^2_{iTm^*l})$, $i = O, A$, $\kappa_{Dl} = I_{\{m^*=12\}}/s^2_{DTm^*l}$, and $\kappa_{Cl} = 1/\hat{\sigma}^2_{\eta l}$. Note that the above posterior means are under the unconstrained model. The adjustment d_l is the difference of the posterior means under the constrained and unconstrained models; that is, $d_l = \hat{\mu}^{(C)}_{Tl} - \hat{\mu}^{(U)}_{Tl}$. Note from (10) that the relative sizes of d_l for $l = 1, \ldots, L$ are proportional to w_{Tl}.

4. Conclusion

In conclusion, the methodology presented allows for yield estimates that are reproducible and objective and produces statistically based error estimates. The model-based yield estimates have been presented to the ASB for consideration during the 2011 and 2012 growing seasons.

A. Spec-region-level parameter estimation via MCMC

A.1 Prior distributions

We place diffuse prior distributions on the parameters. Specifically, let $b_{km^*} \sim$ iid $N(0, \delta_b^2)$ and $\boldsymbol{\beta} \sim N(0, \delta_\beta^2 \boldsymbol{I})$, where $\boldsymbol{I}$ is the identity matrix, for $\delta_b^2 = \delta_\beta^2 = 10^6$ and $\sigma^2_{km^*}, \sigma^2_\eta \sim$ iid IG(A_σ, B_σ), where "IG" refers to the inverse gamma distribution and $A_\sigma = B_\sigma = 0.001$.

A.2 Gibbs sampling algorithm

Let $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_T)'$ and let $\boldsymbol{Z}$ denote a matrix with rows $\boldsymbol{z}_1', \ldots, \boldsymbol{z}_T'$. The following describes a single iteration of the Gibbs sampling algorithm.

1. For $t = 1, \ldots, T$, generate μ_t using $\mu_t|\boldsymbol{\Omega} \sim N(\Delta_{2t}/\Delta_{1t}, 1/\Delta_{1t})$, where $\Delta_{1t} = \sum_{k=O,A} I_{\{t\geq t_k\}}/(\sigma^2_{km^*}+s^2_{ktm^*}) + I_{D(t,m^*)}/s^2_{Dt,12} + 1/\sigma^2_\eta$ and $\Delta_{2t} = \sum_{k=O,A} I_{\{t\geq t_k\}}(y_{ktm^*} - b_{km^*})/(\sigma^2_{km^*} + s^2_{ktm^*}) + I_{D(t,m^*)}y_{Dt,12}/s^2_{Dt,12} + \boldsymbol{z}_t'\boldsymbol{\beta}/\sigma^2_\eta$, with $D(t, m^*) = \{t_D \geq t \geq T-1\}$ or $\{t = T, m^* = 12\}$.

2. For $k = O, A$, generate b_{km^*} according to $b_{km^*}|\boldsymbol{\Omega} \sim N(\Delta_{2,b_{km^*}}/\Delta_{1,b_{km^*}}, 1/\Delta_{1,b_{km^*}})$, where $\Delta_{1,b_{km^*}} = \sum_{t=1}^{T} I_{\{t \geq t_k\}}/(\sigma^2_{km^*}+s^2_{ktm^*})+1/\delta_b^2$ and $\Delta_{2,b_{km^*}} = \sum_{t=1}^{T} I_{\{t \geq t_k\}}(y_{ktm^*} - \mu_t)/(\sigma^2_{km^*} + s^2_{ktm^*})$.

3. Generate $\boldsymbol{\beta}$ using $\boldsymbol{\beta}|\boldsymbol{\Omega} \sim N(\boldsymbol{\Delta}_{1\beta}^{-1}\boldsymbol{\Delta}_{2\beta}, \boldsymbol{\Delta}_{1\beta}^{-1})$, where $\boldsymbol{\Delta}_{1\beta} = \boldsymbol{Z}'\boldsymbol{Z}/\sigma^2_\eta + \delta_\beta^{-2}\boldsymbol{I}$ and $\boldsymbol{\Delta}_{2\beta} = \boldsymbol{Z}'\boldsymbol{\mu}/\sigma^2_\eta$.

4. Generate σ^2_η according to $\sigma^2_\eta|\boldsymbol{\Omega} \sim \text{IG}(T/2 + A_\sigma, (\boldsymbol{\mu} - \boldsymbol{Z\beta})'(\boldsymbol{\mu} - \boldsymbol{Z\beta})/2 + B_\sigma)$.

5. Generate $\sigma^2_{km^*}$ for $k = O, A$ using a Metropolis-Hastings algorithm with a lognormal proposal distribution. That is, we generate the proposed values $\sigma^{2(*)}_{km^*} = \exp(\gamma_k)$ for $k = O, A$, where $\gamma_k \sim N(\log \sigma^{2(j-1)}_{km^*}, \sigma^2_\gamma)$. In the previous, $\sigma^{2(j-1)}_{km^*}$ refers to the value of σ_{km^*} on the previous iteration and $\sigma^2_\gamma = 0.05$. To determine whether the proposed values are accepted, we calculate $R = R_O R_A$, where

$$\begin{aligned} R_k = &\left(\frac{\sigma^{2(*)}_{km^*}}{\sigma^{2(j-1)}_{km^*}}\right)^{-A_\sigma} \exp\left\{-B_\sigma\left[\frac{1}{\sigma^{2(*)}_{km^*}} - \frac{1}{\sigma^{2(j-1)}_{km^*}}\right]\right\} \\ &\times \exp\left\{\sum_{t=1}^{T} I(t \geq t_k)[LL_{k,t}(\boldsymbol{\theta}^{(*)}) - LL_{k,t}(\boldsymbol{\theta}^{(j-1)})]\right\}, \end{aligned} \tag{14}$$

for $k = O, A$, where $LL_{k,t}(\boldsymbol{\theta}) = -(1/2)\log(\sigma^2_{km^*} + s^2_{ktm^*}) - (1/2)(y_{ktm^*} - \mu_t - b_{km^*})/(\sigma^2_{km^*} + s^2_{ktm^*})$ and $\boldsymbol{\theta}^{(*)}$ and $\boldsymbol{\theta}^{(j-1)}$ use $\sigma^{2(*)}_{km^*}$ and $\sigma^{2(j-1)}_{km^*}$, respectively. Then, the proposed values $\sigma^{2(*)}_{km^*}$, $k = O, A$, are accepted with probability $\min(R, 1)$; otherwise, the values $\sigma^{2(j-1)}_{km^*}$ are maintained.

REFERENCES

Caristi, J. (2012) Agricultural Statistics Board performance. RDD Research Report #RDD-yy-xx. Washington, DC: USDA, National Agricultural Statistics Service.

Doraiswamy, P.C., Sinclair, T.R., Hollinger, S., Akhemedov, B., Stern, A., Prueger, J. (2005). Application of MODIS derived parameters for regional yield assessment. Remote Sensing of Environment. 97: 192-202.

Gelman, A., Carlin, J., Stern, H., and Rubin, D. (2003) Bayesian Data Analysis (2nd ed.), London: Chapman & Hall/CRC.

Keller, T. and Olkin I. (2002). Combining correlated unbiased estimators of the mean of a normal distribution. Stanford University Technical Report No. 2002-5.

Nandram, B. and Sayit, H. (2011). A Bayesian analysis of small area probabilities under a constraint. Survey Methodology, 37: 137-152.

N.O.A.A. National Environmental Satellite, Data, and Information Service (NESDIS) (2012). Divisional Data Select. Retrieved September 30, 2012, from http://www7.ncdc.noaa.gov/CDO/CDODivisionalSelect.jsp.

Office of Management and Budget (OMB) (2006). Standards and guidelines for statistical surveys. Washington DC: Government Printing Office.

Tannura, M.A., Irwin, S.H., and Good, D.L. (2008) Weather, technology, and corn and soybean yields in the U.S. corn belt. Marketing and Outlook Research Report 2008-01, Department of Agricultural and Consumer Economics, University of Illinois at Urbana-Champaign.

U.S.D.A. National Agricultural Statistics Service (NASS) (2006). The yield forecasting program of NASS. Statistical Methods Branch (SMB) Report #SMB 06-01. Washington DC: USDA.

U.S.D.A. National Agricultural Statistics Service (NASS) (2009). 2007 Census of Agriculture, United States Summary and State Data. Volume 1: Geographic Area Series. Part 51. Report AC-07-A-51. Washington DC: USDA.

U.S.D.A. National Agricultural Statistics Service (NASS) (2012a). Crop Production. Washington DC: USDA.

U.S.D.A. National Agricultural Statistics Service (NASS) (2012b). Quick Stats. In USDA NASS. Retrieved September 30, 2012, from http://www.studygs.net/citation.htm.

U.S.D.A. World Agricultural Outlook Board (WAOB) (2012). World Agricultural Supply and Demand Estimates. WASDE-510. Washington DC: USDA.

Vogel, F.A. and Bange, G.A. (1999). Understanding USDA Crop Forecasts. Miscellaneous Publication No. 1554. Washington, DC: USDA.

Wang, J.C., Holan, S.H., Nandram, B., Barboza, W., Toto, C., and Anderson, E. (2011) A Bayesian approach to estimating agricultural yield based on multiple repeated surveys. Journal of Agricultural, Biological, and Environmental Statistics, 17(1): 84-106.

Warren, F.B. (1985). Evaluation of the 1983 Corn Objective Yield Validation Survey. SRS Staff Report No. AGES850109. Washington DC: USDA, Statistical Reporting Service.