Small Area Modeling of County Estimates for Corn and Soybean Yields in the U.S.

Matthew Williams

Research and Development Division
National Agricultural Statistics Service
United States Department of Agriculture
1400 Independence Avenue, SW
Washington, DC 20250

Abstract

The National Agricultural Statistics Service (NASS) conducts the County Agricultural Production Survey (CAPS) to establish county-level acreage and production. Major changes over previous programs allow for results to be combined with the year-end production surveys with the goal of improving accuracy and defensibility. The goal of this research is to evaluate the use of a class of modeling techniques known as *Small Area Models* to add value to the county estimation process at NASS. We apply several of these methods to estimate corn and soybean yields for 3 of the pilot states from 2010 and for 10 of the production states in 2011 and 2012. Comparing performance of estimators to the final Board yield, the models show little improvement over the survey indications. Possible explanations include model misspecification, poor or noisy covariate information, the shrinkage nature of the model estimates, and/or the use of the final Board yields as the standard for comparison. It is recommended that we use results of the 2012 Census of Agriculture as a standard to compare the model estimates, the survey indications, and the 2012 final Board values. It is also recommended that related research continue for benchmarking and utilizing alternative covariate and variance structures to improve the flexibility of the models.

1. Introduction

In 2009-2010, the National Agricultural Statistics Service (NASS) initiated the county estimates pilot study in 5 states as a new approach for selecting and processing county estimates that was consistent with procedures used for the Agricultural Survey programs. Because county estimates data are merged with year-end survey data (September Crops/Stocks and December Crops/Stocks), the county estimates survey processes should be as consistent as possible. These year-end surveys have consistent sample designs, data collection modes, edit specifications and summary routines. To improve statistical defensibility of NASS county estimates, the merged data used to set county level estimates should be as consistent as possible between the year-end surveys and the data from the County Agricultural Production Survey (CAPS). The 2009-2010 pilot study produced indications with variances and CVs at the county level for the 5 states in the study. For 2011 the probability design was expanded to cover all States.

The CAPS results in survey indications that are incorporated into the Database Integrated County Estimates (DICE) system along with administrative data such as Farm Service Agency (FSA) planted acres and remote sensing indications. Statisticians within each state establish state recommendations and then HQ staff establish estimates via a Board review process. While the new CAPS survey produces indications and measures of uncertainty (standard errors, or equivalently, coefficients of variation), the current review process establishes estimates without a quantitative measure of the associated uncertainty.

The goal of this research is to evaluate the use of a class of modeling techniques known as *Small Area Models* to add value to the county estimation process within NASS. Potential gains include increased automation and efficiency, and the added functionality of producing measures of uncertainty associated with those estimates.

2. Small Area Models

Estimates for small domains or areas are classified into *direct* and *indirect* or *design-based* and *model-based* (Rao, 2003). In general, *direct* estimates only use survey and covariate information from individual areas to create estimates (or indications) for those areas. In contrast, *indirect* estimators may use information from the past or from other areas in the construction of estimates. The term *design-based* tends to be used interchangeably with *direct* estimates, whereas the term *model-based* is typically associated with *indirect* estimates. Of course these labels are somewhat deceiving, since both direct and indirect estimates may utilize *both* model *and* survey design features such as sampling weights. The methods described below are considered *model-based* and rely on nested error (mixed model) regression to combine survey and covariate information. These models include a correlation between small areas, thereby pooling information across those areas. This is an example of what is typically referred to as "borrowing strength", a phrase ubiquitous in the literature.

2.1 Shrinkage Estimators

The classic small area models for means or totals are related to a class of estimators known as "shrinkage estimators". To better understand these estimators, we first look at the so-called "Paradox" of Charles Stein:

Suppose we have a collection of county indications y which we believe have a true expected value $E(y) = \theta$. Then one popular criterion for whether or not a particular set of estimates $\hat{\theta}$ are "good" is to evaluate the Mean Square Error (MSE): $E\left[\left(\theta - \hat{\theta}\right)^2\right] = E\left[\left(\theta - \hat{\theta}\right)\right]^2 + E\left[\left(E[\hat{\theta}] - \hat{\theta}\right)^2\right]$ which can be decomposed into two components defining the bias of the estimator $\hat{\theta}$ and its variance. In order to get a "good" set of estimates $\hat{\theta}$, we want to reduce the bias or the variance of the estimator or both. We tend to assume that survey indications are unbiased, in which case, a "good" set of survey indications would be one that has small variance.

Stein's Paradox takes aim at the least squares estimator, which is an unbiased estimator that minimizes the variance for all unbiased estimators. The least squares estimator is not unlike our

survey indications. Stein's results show that for a set of estimates of size 3 or greater, the regular least squares estimator is inadmissible. It turns out we can construct estimators that lead to total MSE (MSE summed over all estimates) which are uniformly lower than the least squares estimator over all possible values for the true mean θ. The basic idea behind this property is that we consider estimates that introduce a small amount of bias but lead to a large reduction in variance. The net result is a decrease in total MSE.

Stein's Estimators often take the form of what is called a shrinkage estimate:

$$\hat{\theta}_{sh} = \bar{y} + c(\hat{\theta}_{LS} - \bar{y})$$

where c is a collection of scaling factors. This is called a shrinkage estimator because values are "shrunk" away from the Least Squares solution $\hat{\theta}_{LS}$ toward the mean $\bar{y}$. We can rework the notation to form a so-called "composite" estimate, which is a weighted average of two separate estimates:

$$\hat{\theta}_c = \mathrm{c}\,\hat{\theta}_A + (1 - \mathrm{c})\hat{\theta}_B$$

The scaling factors c can be chosen in several ways. There is a *Scientific American* article (Efron & Morris, 1977) that gives a nice discussion on the Stein estimator along with several examples. The small area models which we consider here have the survey indications as the first estimator $\hat{\theta}_A$ and a linear regression model as the second estimator $\hat{\theta}_B$. The scaling factors c are determined by the relative variance between these two sets of estimators.

2.2 County-level and Record-level Models

Small area models utilize covariate information to construct a modeled value and then create a weighted composite of that value and the survey indication (See previous section). Covariate information may be available at the county level or the record level. Models that use only county-level covariates and survey indications are called *area-level* models. Models that use record-level covariates and responses are called *unit-level* models. Unit-level models can also incorporate area-level covariates, but typically do not use area-level survey indications. See the Appendices for more details.

2.2.1 County-level Models

County-level (i.e., area-level) models use survey indications and standard errors directly. These are combined with a linear model based on covariate information at the county level. The scaling factor c is determined by standard errors of the survey indication and a model-based estimate of between-county variance.

The main benefits of county-level models are that they are simpler than the record-level models, require only county-level covariates, and are "design-consistent". If the sample size (and the population) of the survey increases to infinity, the county-level model will give the same results as the survey indications. From our perspective, a more practical benefit is that the survey indication and its estimated standard error are used directly so that all the sampling weight corrections and adjustments from summarization have been incorporated.

Potential drawbacks for the county-level models are that we ignore record level covariates and responses. These might provide more information and let us flag or re-weight unusual observations. Of course for NASS, this may have already been done during summarization.

2.2.2 Record-level Models

Record-level (i.e., unit–level) models use survey responses at the record level. They use weights (either sample size based or from sampling weights) and record- and county-level covariates to model the county-level mean and the scatter of the records about it. The scaling factor c is determined by model estimates of between-county variances and the residual variance of the model.

The main benefits of the record-level models are that they can use all available covariates whether at the record or county level. They model the distribution of records about the county mean allowing for the potential incorporation of robustness measures to down-weight unusual observations (although in practice this is very limited).

Limitations of using record-level models are that all records used to generate the survey indication are needed. In order to use a record-level covariate, not only are the covariate values needed for every record used, county-level totals for this covariate are also needed. In other words, we need some covariate information about records that were not even sampled, at a minimum, the totals for this group. Record-level models are not "design-consistent": if sample and population sizes increase to infinity, the model estimates may not converge toward the survey indication. For our purposes, a more important issue is how to incorporate the survey weights in a meaningful way. For CAPS, NASS adjusts the original sampling weights and uses replicate weights to estimate standard errors via a jackknife procedure. Deciding which weights to use and how to use them is not simple. Furthermore, not all records are used or at least used in the same way. Tracking down how the summary system treats different classes of records is a daunting task, and incorporating different uses of records is not mentioned in the literature for small area models.

2.3 Parameter Estimation

If the parameters in the small area models were known, the model estimates would follow naturally and look like the shrinkage or composite estimators discussed above. Most often, parameters need to be estimated. In the context of small area models, there are two related approaches to estimation which we will call Empirical Bayes (EB) and Hierarchical Bayes (HB). Neither name is particularly informative, but both are prevalent in the literature (Rao, 2003).

In the context of our small area models, EB approaches construct a probability distribution for the data and maximize the likelihood (or a restricted likelihood) to obtain parameter estimates. These estimated parameters are then "plugged" into expressions to obtain county estimates and asymptotic (approximate) standard errors. The HB approach takes the same probability model for the data as the EB approach. In addition, it constructs a prior probability distribution for the parameters. It then takes the two sets of distributions and combines them using Bayes' Theorem. The result is an updated, more specific posterior distribution for the unknown parameters.

Together, these distributions are used to produce estimates and measures of variability for each county.

In practice, both the EB and HB methods often agree for simple models and the choice of one or the other is based on professional training, personal preference and experience, or implementation issues such as computation time and availability of software. The main reasons that both names are deceiving are that (1) Both models are hierarchical in nature and (2) Empirical Bayes is really a likelihood or "frequentist" approach rather than what is typically called "Bayesian".

3. Pilot Survey (2010)

Of the five states in the pilot study, three were selected for modeling. The other two states had three or fewer counties with corn and soybean indications, and thus were excluded. Using the data from these states, we compared a variety of models (area-level, unit-level with and without sampling weights) using both parameter estimation approaches (EB and HB).

3.1 Covariates

Only county-level covariates were available. These included the 2009 state statistician recommended ("stat") yield and survey indications for yield, FSA planted acreage, the National Commodity Crop Productivity Index (NCCPI), and normalized difference vegetation index (NDVI) based values. The NCCPI is an index produced by the Natural Resource Conservation Service (NRCS) which rates land by its potential for production based on soil and other physical characteristics. There are some commodity specific indices, but we consider only the general one. In other words, we have the same covariate for corn and for soybeans for any county which produces both crops. For more information about the NCCPI, see the user guide (NRCS, 2008).

The NDVI is an index derived from remote sensing data and is typically on a scale from 0 to 1. It is often purported as measuring the "greenness" of an area. Tracking this "greenness" for areas where the commodity is known to grow can give an indication of expected yield. From this time series of NDVI values (16 days apart), we derive three related covariates: the peak (pNDVI), the trimmed mean (tmNDVI), and the total excess or area over 0.7 (a7NDVI). The trimmed mean is for a period from June 10th to Aug. 29th with the first and last values down-weighted. For more information and references for NDVI, the U.S. Geological Survey (USGS) has a helpful webpage http://ivm.cr.usgs.gov/whatndvi.php (USGS, 2010).

While several sets of these covariates (2009 stat yield and survey indications; peak, trimmed mean and excess NDVI) are correlated, we have included them all in the hopes of getting the best fit. If the fit is deemed adequate, then we should consider some variable or model selection method or criteria to counteract possible issues arising from multicollinearity.

3.2 Results

The survey indications for 2010 were used as the response input into the model. Both the original indications and the model estimates from each combination of model and parameter estimation

approach were compared to the final Board yields for 2010. These final values were either published or suppressed due to confidentiality or other publication rules.

We determine which set of estimates are the "best" *quantitatively* by comparing overall bias (mean of differences), mean square error (mean of squared differences), and mean absolute deviation (mean of absolute differences) and *qualitatively* by comparing plots of the modeled values vs. the final Board. The quantitative metrics give an overall or global assessment of the estimates, whereas the plots give a qualitative county by county perspective of which values are driving the differences between metrics. See the Appendix for some examples.

Overall, the 2010 survey indications for yield were the best predictors for the final 2010 Board yield for corn and soybeans. For one state, there was minor improvement of the models over the survey indications in terms of the quantitative metrics, but no noticeable improvement qualitatively in the plots. The second best set of estimates came from the county-level models. In addition, there was strong agreement between the estimates from the EB and the HB approaches. While this is expected, this agreement was much tighter than for the case of the record-level models.

Among the various record-level models, those which incorporated the survey weights showed the most variability. It should be noted that the lack of agreement between the EB and HB approaches and the noticeable variability in the record-level methods was not observed in our simulation studies used in developing the programming code. This suggests possible misspecification in the models or other issues.

Finally, as part of the HB estimation process, various diagnostics are used to determine if the procedure has converged and is behaving as expected. Although the diagnostics were generally acceptable for all the models, the diagnostics for the county-level models indicated superior mixing of the Markov chain. This suggests that the county-level model is reasonably specified and can run with fewer iterations than the record-level ones.

4. Full Production Survey (2011 & 2012)

For 2011-12, the probability design was implemented in all states conducting the CAPS. Of these states, 10 were selected to be modeled. These states were chosen to represent large producers of corn and soybeans, and to have some continuity with the 2010 pilot states. Based on the results from modeling the 2010 pilot states, only county-level models using both EB and HB methods were used.

4.1 Covariates

We only consider county-level covariates (complete record-level covariates were not available). These included NCCPI, and the three NDVI-based values (tmNDVI, pNDVI, and a7NDVI). In addition, a remote sensing derived indication was provided by Spatial Analysis Research Section (USDA/NASS) for 9 of the 10 states. This is also highly correlated with the NDVI values (see previous section). For 2011, models were run with and without the remote sensing indication.

For 2012, all models (9 of 10 states) were run using the remote sensing indication. FSA planted acreage was left out, since there appeared to be no relationship in the 2010 data. Previous year's values were also excluded since 8 of the 10 states were not in the probability sample in 2010 and, more importantly, because the 2010 models showed little effect from including previous year (2009) indications.

4.2 Results

The survey indications for 2011 (and 2012) were used as the response input into the model. Both the original indications and the model estimates for each parameter estimation approach (EB and HB) were compared to the final Board yields for 2011 (and 2012). These final values were either published or suppressed due to confidentiality or other publication rules.

Following the same criteria as for the 2010 models, we determined which set of estimates were the "best" *quantitatively* by comparing overall bias (mean of differences), mean square error (mean of squared differences), and mean absolute deviation (mean of absolute differences), and *qualitatively* by comparing plots of the modeled values vs. the final Board. See the Appendix for some examples.

For the 10 states, no appreciable gains were made by using the models over the original survey indications. For four commodity-state-year combinations, modest improvements in two or more quantitative metrics were made. Qualitatively, there was not much improvement. Most of the gains could be explained by a change to a single county. For a half-dozen other combinations, the models and the survey indications more or less "break even" based on the metrics and the plots. For the other commodity-state-year combinations (about 30), there was not much quantitative or qualitative evidence to suggest that models were doing anything more than adding noise to the process.

The inclusion and exclusion of the remote sensing indication as a covariate had little bearing on the results for 2011, so this comparison was not repeated for 2012. Estimates between the two sets of models were slightly different, but the overall performance of the models relative to the survey indications was essentially the same. The same can be said for the EB and HB methods with the agreement between them being even tighter. This is expected, because the EB and HB methods are simply two different approaches to estimating the same models, rather than two separate models.

5. Discussion and Conclusions

Using standard (or even more complex) small area estimation models based on mixed models (shrinkage estimates) seems to be problematic for county level yield estimates. From applying these models to corn and soybean yield for the 2010 pilot states and 10 states from 2011-12, the results generally indicate no real improvement over the survey indications.

One potential cause for the ineffectiveness of these models is the direction of causality. By comparing the model results and survey indications to the final Board values and treating those as the standard or 'truth', we are essentially assuming that the survey indications were generated

by the Board and uncertainties arose due to the sampling design and survey process. Since this is obviously not the case, the adjustments that the Board or commodity statisticians make to account for weather, growing conditions, and other factors should be incorporated somehow.

We see evidence for such adjustments when we compare the survey indications to the final Board values. Often, the values are mostly unchanged except for a few large changes. Some systematic changes are present, suggesting an effort to benchmark or "ratio-up" counties to hit state or district targets. In contrast, the small area models tend to change all the counties at least a little. This spreading out of the change can be a good property in many applications, but it seems to add noise to the 'good' counties that should not be adjusted from their survey indication.

Another likely cause for the models' poor performance is that we assume the covariates have strong linear relationships with the 'true' yield. Using the survey indications or the Board values, we can see that all the covariates are relatively noisy and weakly related to yield. Without a stronger relationship, we would not expect the model to improve much upon the survey indications. This relationship could even take a nonlinear form as long as the noise was lower (see related research below).

Finally, it is clear that the final Board yields are the best available yields within a certain time frame after the growing season, but it is unclear how CAPS as whole can be evaluated. If other metrics or administrative data are available much later, we should try to incorporate this into our analysis of model performance. Fortunately, the 2012 growing season will have corresponding values from the 2012 Census of Agriculture to use as a comparison when they are published in 2014.

6. Related Research

While the above research is on-going, there are some related projects both realized and potential that should add value to our process. However, none of these methods "solve" the issues addressed above.

6.1 Benchmarking

Since yield is a ratio of production per harvested area, adjustments must be made to ensure that county yield, production, and harvested acreage agree in aggregation with the official state estimates. While the default approach (benchmarking acreage first, then using acreage and yield to construct production, and finally benchmarking production) is reasonable, there are other methods that can be used. A recent paper (Williams & Berg, 2013) describes several methods and attempts to create a framework that allows for an analyst to make adjustments to these automated procedures. In general, benchmarking can be significantly more complex than simply applying ratio adjustment across all counties.

6.2 Nonlinear Relationships

Although weak relationships between covariates and the survey indications are a potential source of problems (see above), strong relationships (high signal to noise) between covariates and

indications don't need to be in a linear form. The techniques of general additive models have already been applied to incorporate nonlinear relationships for the record-level model (Opsomer, Claeskens, Ranalli, Kauermann, & Breidt, 2008). Similar techniques may also be used for the county-level models. In other words, in the search for more useful covariates, we need not restrict our attention to only those which have a strictly linear relationship with the survey indications.

6.3 Alternative Distribution of Error Terms

One extension of the record-level model involves adjusting the modeled variance associated with individual records. In particular, some records are allowed to have higher variance. The net effect is that potential "outlier" responses are down-weighted, resulting in estimates that are less sensitive or more "robust" to unusual observations. Dissertation work from the Joint Program in Survey Methodology (Gershunskaya & Lahiri, 2011) has made strides to implement such a method using a scaled mixture of normal distributions with parameter estimation via an EM algorithm.

References

Efron, B., & Morris, C. (1977). Stein's Paradox in Statistics. *Scientific American , 236*, 119-127.

Gershunskaya, J., & Lahiri, P. (2011). Robust Small Area Estimation Using a Mixture Model. *Proceedings of the 58th World Statistics Congress.* Dublin: The International Statistical Institute Proceedings of the 58th World Statistics Congress.

NRCS. (2008). *User Guide National Commodity Crop Productivity Index (NCCPI) Version 1.0.* Retrieved 2012, from ftp://ftp-fc.sc.egov.usda.gov/NSSC/NCCPI/NCCPI_user_guide.pdf

Opsomer, J. D., Claeskens, G., Ranalli, M. G., Kauermann, G., & Breidt, F. J. (2008). Non-parametric small area estimation using penalized spline regression. *Journal of the Royal Statistical Society Series B , 70*, 265-286.

Prasad, N., & Rao, J. (1999). On robust small area estimation using a simple random effects model. *Survey Methodology , 25*, 67-72.

Rao, J. (2003). *Small Area Estimation.* New Jersey: John Wiley & Sons.

USGS. (2010). *Greenness of the Conterminous U.S.* Retrieved 2012, from http://ivm.cr.usgs.gov/whatndvi.php

Williams, M., & Berg, E. (2013). Incorporating user input into optimal constraining procedures for survey estimates. *Journal of Official Statistics* , 375-396.

You, Y., & Rao, J. (2002). A pseudo-empirical best linear unbiased prediction approach to small area estimation using survey weights. *Canadian Journal of Statistics , 30*, 431-439.

You, Y., & Rao, J. (2003). On robust small area estimation using a simple random effects model. *Journal of Statistical Planning and Inference , 111*, 197-208.

A. Area Level Model Details

If we have direct estimates $\hat{\theta}_i$ from the survey indications and the associated standard errors ψ_i (or estimates of them) for all counties $i = 1, \dots, m$ as well as covariates at the county level, then we can construct an area level model. From Rao (2003)

$$\hat{\theta}_i = z_i^T \beta + b_i v_i + e_i$$

where z_i is a covariate vector for county i, $v_i \sim N(0, \sigma_v^2)$, and $e_i \sim N(0, \psi_i)$. Given σ_v^2, the best linear unbiased predictor (BLUP) for the county mean is:

$$\tilde{\theta}_i^H = \gamma_i \hat{\theta}_i + (1 - \gamma_i) z_i^T \tilde{\beta}$$

with $\gamma_i = \sigma_v^2 b_i^2 / (\psi_i + \sigma_v^2 b_i^2)$ and

$$\tilde{\beta}(\sigma_v^2) = \left[\sum_{i=1}^{m} z_i z_i^T / (\psi_i + \sigma_v^2 b_i^2) \right]^{-1} \left[\sum_{i=1}^{m} z_i \hat{\theta}_i^T / (\psi_i + \sigma_v^2 b_i^2) \right]$$

When σ_v^2 is unknown, we use EB and HB methods to estimate σ_v^2, β, and $\tilde{\theta}_i^H$ simultaneously (Rao, 2003). For our models we assume that $b_i = 1$ for all counties.

B. Unit Level Model Details

If we have direct responses from the survey records y_{ij} for counties $i = 1, \dots, m$ and records $j = 1, \dots, n_i$ and covariates $\boldsymbol{x}_{ij}$ for the records, then we can use a unit level model. From Rao (2003)

$$\boldsymbol{y}_i = \boldsymbol{X}_i \beta + v_i \mathbf{1}_{n_i} + \boldsymbol{e}_i$$

with $v_i \sim N(0, \sigma_v^2)$ and $\boldsymbol{e}_i \sim N(0, \sigma_e^2 \boldsymbol{I}_{n_i})$. The BLUP estimator for county mean μ_i is then

$$\tilde{\mu}_i^H = \gamma_i \left[\bar{y}_i + (\bar{X}_i - \bar{x}_i)^T \tilde{\beta} \right] + (1 - \gamma_i) \bar{X}_i^T \tilde{\beta}$$

where $\bar{X}_i$ is the average of the covariate values over the entire population for county i and $\bar{x}_i$ is the average covariate value of the sample from county i.

When the two error variances σ_v^2 and σ_e^2 are given, the estimate $\tilde{\beta}$ is the best linear unbiased estimator (BLUE) of β:

$$\tilde{\beta} | \sigma_v^2, \sigma_e^2 = \left[\sum_i \boldsymbol{X}_i^T \boldsymbol{V}_i \boldsymbol{X}_i \right]^{-1} \left[\sum_i \boldsymbol{X}_i^T \boldsymbol{V}_i \boldsymbol{y}_i \right]$$

with $\boldsymbol{V}_i = \sigma_e^2 \boldsymbol{I}_{n_i} + \sigma_v^2 \mathbf{1}_{n_i} \mathbf{1}_{n_i}^T$.

When σ_v^2 and σ_e^2 are unknown, we use EB and HB methods to estimate σ_v^2, σ_e^2, β, and $\tilde{\mu}_i^H$ simultaneously (Rao, 2003).

B.1 Sampling Weights

If unit level data are available, we often have their sampling weights w_{ij} which are usually the inverse of the selection probabilities π_{ij}. It may be more prudent to use final calibrated weights to get results consistent with the direct estimators $(\bar{y}_{iw} = \hat{\theta}_i)$. For CAPS, replication weights are also used to estimate variance via a delete-a-group jackknife procedure. These variations surrounding sampling weights for small area models are not present in the literature. The main question in the literature is whether to use equal weights or sampling weights.

Taking the weighted averages of the responses and the covariates: $\bar{y}_{iw} = \sum_j w_{ij} y_{ij} / \sum_j w_{ij}$ and $\bar{\boldsymbol{x}}_{iw} = \sum_j w_{ij} \boldsymbol{x}_{ij} / \sum_j w_{ij}$. Then taking the weighted sum of the unit level model we get the aggregated area level model:

$$\bar{y}_{iw} = \bar{\boldsymbol{x}}_{iw}^T \beta + v_i + \bar{e}_{iw}$$

with $\bar{e}_{iw} \sim N(0, \delta_i)$ where $\delta_i = \sigma_e^2 \sum_j \left(w_{ij}^2\right) / \left(\sum_j w_{ij}\right)^2$.

Given values for σ_e^2 and σ_v^2, we can estimate β and then $\tilde{\mu}_i^H$ using $\gamma_{iw} = \sigma_v^2 / (\sigma_v^2 + \delta_i)$. From the literature, there are at least three ways to estimate β:

1. Take $\tilde{\beta}$ from the unit level model directly. Estimate $\tilde{\mu}_i^H$ with $\gamma_i = \gamma_{iw}$.
2. Use the conditional predictor of v_i from above: $\tilde{v}_{iw}(\beta, \sigma_e^2, \sigma_v^2) = \gamma_{iw}(\bar{y}_{iw} - \bar{\boldsymbol{x}}_{iw}^T \beta)$.
    a. Then solve the weighted estimating equations for β:
    $$\sum_i \sum_j w_{ij} \boldsymbol{x}_{ij} \left[y_{ij} - \boldsymbol{x}_{ij}^T \beta - \tilde{v}_{iw}(\beta, \sigma_e^2, \sigma_v^2) \right] = 0$$
    b. The sum form of $\tilde{\beta}_w$:
    $$\tilde{\beta}_w = \left[\sum_i \sum_j w_{ij} \boldsymbol{x}_{ij} \left(\boldsymbol{x}_{ij} - \gamma_{ij} \bar{\boldsymbol{x}}_{iw} \right)^T \right]^{-1} \left[\sum_i \sum_j w_{ij} \left(\boldsymbol{x}_{ij} - \gamma_{ij} \bar{\boldsymbol{x}}_{iw} \right) y_{ij} \right]$$
    c. The matrix form of $\tilde{\beta}_w$: Let $z_{ij} = w_{ij} \left(\boldsymbol{x}_{ij} - \gamma_{ij} \bar{\boldsymbol{x}}_{iw} \right)$. Then $\tilde{\beta}_w = (\boldsymbol{X}^T \boldsymbol{Z})^{-1} \boldsymbol{Z}^T \boldsymbol{y}$.
3. Use the area level-like relationship from above to get:
$$\tilde{\beta}_w = [\boldsymbol{X}_w^T \boldsymbol{\Gamma}_w \boldsymbol{X}_w]^{-1} \boldsymbol{X}_w^T \boldsymbol{\Gamma}_w \boldsymbol{y}_w.$$

Of course we do not know σ_e^2 and σ_v^2, so we use the unit level model to provide estimates either through EB or HB methods. The resulting estimators for $\tilde{\mu}_i^H$ are known as Pseudo EB (Pseudo EBLUP) and Pseudo HB estimators. Both methods have associated estimators for variances as well. More details for these methods are available in the literature. Rao (2003) introduces Method 2 for EB and HB. More detail can be found in You & Rao (2002) and You & Rao (2003) for EB and HB respectively. Method 3 is introduced in Prasad & Rao (1999) for EB and evaluated for HB in You & Rao (2003). Method 1 is examined in You & Rao (2003) for HB.

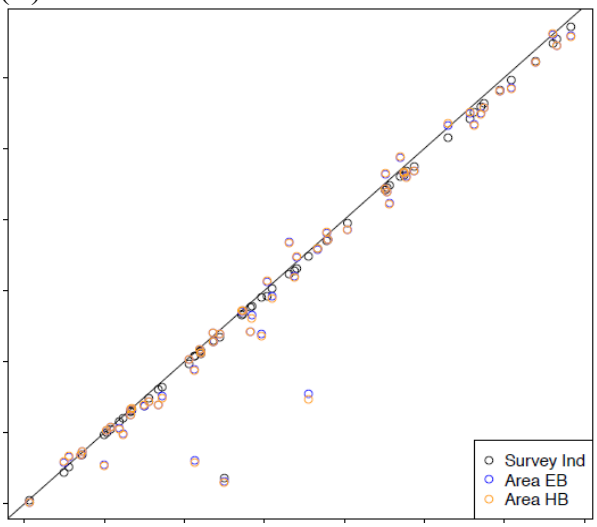
All three methods for using sampling weights are considered 'design consistent', meaning that as sample size in each small area grows, the model estimates $\tilde{\mu}_i^H$ should converge to the direct survey estimates $\hat{\theta}_i$ (assuming that we use the correct weights). This property (which is asymptotic) also holds for the regular area level model, but not the regular unit level model. Methods 1 and 3 are simpler to implement. Method 2 should have an advantage in terms of smaller estimated variance (more precision) for estimates. If sampling weights are calibrated to total population size, Method 2 produces estimates that are self-benchmarked to the direct estimate for the total population. This may be a desirable property since any benchmarking would alter the variance estimates for Methods 1 and 3. For CAPS we are not internally benchmarking, but instead have the published or final Board values for states as the target, so all estimates would have to be benchmarked. While Method 2 appears to be superior when the models are correctly specified, it may be more sensitive to misspecification of the weights w_{ij} and it is computationally more intensive since it uses all the record level observations.

C. Examples of Metrics

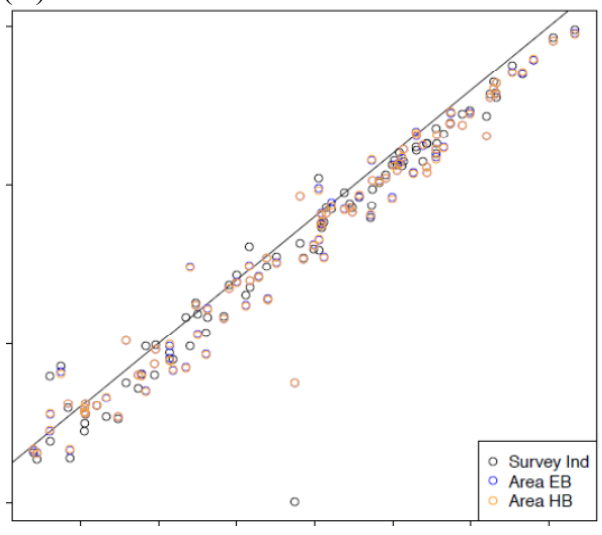
We include some examples of quantitative (Table) and qualitative (Figures) evaluation of estimates. Values in the table are with respect to the final official estimates. Performance is scaled relative to the corresponding metric of the survey indication. Figures plot model and survey values (y-axis) vs. official published estimates (x-axis). (A) Model estimates (EB and HB) show modest gains with slight reductions in bias and RMSE. Most of this can be attributed to gains for a single county, with the other counties receiving more noise. (B) Model estimates "break even" with survey indications showing small gains in RMSE, but showing losses in MAD and Bias. Plots display no noticeable gain, with the farthest points from the line being consistent between the survey indication and the model estimates. (C) No gains from the models are evident. The metrics and plots suggest that the process has increased the noise.

Estimate	Bias	RMSE	MAD
Survey (A,B,C)	1.00	1.00	1.00
EB (A)	0.94	0.81	0.99
HB (A)	0.96	0.82	1.00
EB (B)	-3.25	0.93	1.47
HB (B)	-3.38	0.96	1.56
EB (C)	1.65	1.61	1.94
HB (C)	1.68	1.65	2.02

(B)



(A)



(C)