Mitigating Standard Errors of County-Level Survey Estimates When Data are Sparse

Valbona Bejleri[1], Nathan Cruze[2], Andreea L. Erciulescu[3], Habtamu Benecha[4], Balgobin Nandram[5]

[1]USDA National Agricultural Statistics Service, 1400 Independence Ave., Washington, DC 20250

[2]USDA National Agricultural Statistics Service, 1400 Independence Ave., Washington, DC 20250

[3]National Institute of Statistical Sciences

[4]USDA National Agricultural Statistics Service, 1400 Independence Ave., Washington, DC 20250

[5]Worcester Polytechnic Institute and USDA National Agricultural Statistics Service

Abstract

The USDA National Agricultural Statistics Service's (NASS's) official statistics at the county level are composites of survey and non-survey data that are manually benchmarked to state and national official estimates. NASS is currently developing Bayesian hierarchical models as an alternative to produce county official statistics using survey summaries and auxiliary data as covariates. The modeled county estimates are linear combinations of survey summaries and auxiliary data, with coefficients depending on the standard errors of direct survey estimates. With this approach, the auxiliary data are not used to produce the final model estimate when the standard error of the direct survey estimate is zero. In this paper, it is shown how to mitigate estimated standard errors of zero. The relationship between the direct survey estimates and their standard errors is modeled, if a relationship between the two is present. Exploratory data analysis is conducted and a data driven distribution-based technique using bootstrapping is proposed for cases where the relationship between estimates and their standard errors cannot be modeled well. An illustration of the method using NASS's County Agricultural Production Survey data is presented.

Key Words: Agricultural Survey, Bootstrap, Official Estimates, Small Area Estimation, Zero Variances

1. Introduction

The USDA's National Agricultural Statistics Service (NASS) conducts the County Agricultural Production Survey (CAPS) to produce end-of-year estimates of planted acreage (P), harvested acreage (H), production (G), and yield (Y), where yield is defined as the ratio of G to H, for dozens of small grains and row crops, at the county and district domains. NASS's current county official statistics are constructed by consensus of the Agricultural Statistics Board (ASB) from direct expansions of totals and ratio estimators.

They are manually benchmarked composite of CAPS and other non-survey sources of information. Sampling variances for totals are estimated using a delete-a-group Jackknife and sampling variances for yield are estimated using a second order Taylor series approximation for the ratio (Kott, 1990). NASS is currently testing Bayesian hierarchical models as an alternative for the production of county official statistics, using survey summaries and auxiliary data as covariates. Findings of an external review recommend NASS transition to a system of model-based official statistics (National Academies, 2017).

Modeling approaches to survey estimates were introduced decades ago; however, there is an increased interest in the recent years especially through small area research. The literature suggests that estimates produced from modeling of small area population totals are somewhat more accurate when compared to symptomatic accounting techniques (Ericksen, 1974 b; O'Hare, 1976; Purcell and Kish, 1979). However, models are used to improve the precision of direct estimation even when data are available for every domain (Tzavidis et al., 2018). A combined synthetic-regression method considers the synthetic estimate as a covariate for the regression model, allowing a suitable combination of a biased, low variance synthetic estimate and high variance direct estimate (Nicholls, 1977).

The choice and accuracy of the existing techniques for small domain estimation are dictated by the level and quality of the available data, both response and covariates (Purcell and Kish, 1979). When the synthetic estimate is not available for a sample of small areas, a modeling approach becomes problematic.

NASS county estimates are linear combinations of direct survey summaries and auxiliary data, with coefficients depending on the standard errors (SEs) of direct estimates (DEs) from the survey. With this approach, counties with zero survey estimated variance (SE^2) are not included in the set of modeled counties; i.e., the auxiliary data is not used to produce the final model estimate. A hierarchical lognormal model for the survey variances on the survey estimates for corn harvested acreage is developed in Erciulescu et al. (2018) to mitigate variances strongly related to direct estimates. The parameters of the model are estimated using the subset of sampled data with estimates available (positive) for both quantities. A similar approach may be used to mitigate zero variances. The lognormal model assumption holds for production, but not for yield. Furthermore, exploratory data analyses indicate that there is no easier modeled relationship present between direct estimates of yield, i.e., yield estimate and its variance/standard error.

In this paper, it is shown how to mitigate zero estimated SE of DE of yield at the county level through exploring alternatives other than modeling. All approaches are illustrated using NASS survey data from CAPS. Sample data consist of states that differ by commodity. Concentrating on corn commodity, there were 37 states sampled for corn in 2016. All methods approximate (and replace where applicable) survey variances less than 1 bu/acre. The threshold may differ for other commodities. Throughout the paper, the terms survey variances and (squared) standard errors are used interchangeably. The paper is structured as follows. The problem is set up in Section 2, where a brief description of the subarea level model applied to CAPS summary at NASS is given. The relationship between the DE for total planted area, total harvested area, production and yield for corn and their estimated survey variances is investigated in Section 3. The relationship between the DEs and their SEs is modeled, whenever a "good" relationship between the two is present. A Taylor series approximation and a data driven technique for approximating the distribution of SE of yield is presented in Section 4. Data suggest that the SE of yield for the sampled counties follows a chi-square distribution. The auxiliary variables used as covariates in the

subarea model are further explored to identify the ones that are related to SE of yield. The existing association structure between these variables within the data space is used when estimating the assumed chi-square distribution of the SE of yield. The method of moments and a bootstrap sampling approach are used to estimate the degrees of freedom of the chi-square distribution. Then, all SEs less than 1 bu/acre are replaced with values drawn from the upper-tail of the distribution, i.e., 75th percentile. The final model based county estimates are compared and results are also presented in Section 4. Concluding remarks are given in Section 5.

2. Problem Setup

2.1 County Agricultural Production Survey

The 2016 CAPS sample consists of 37 states comprised of 2881 counties for corn. From these counties, 2467 have positive planted acreage, 2361 counties have positive harvested acreage and 2329 counties have positive yield/production for corn. The subarea level model currently being tested at NASS using CAPS summary considers the DE as the response and assumes that the variance of the DE is available and positive in order to produce the final modeled county estimate. However, due to item level nonresponse, there is a sparseness in the reported data, and CAPS survey summaries could not be produced for every county. This affects yield more than total planted and harvested acreage, and production.

In this paper, the relationships of survey summaries from CAPS for planted and harvested acreage, production, and yield for corn are investigated. In what follows, the CAPS subarea level model applied to at NASS and the issues arising with CAPS summaries when trying to model all counties in the sample are discussed. Finally, alternative solutions to overcome these issues are presented and compared.

First, for county j of district i, denote the DE and SE of total harvested acreage by $(\hat{\theta}_{ijH}, \hat{\sigma}_{ijH})$, of total planted acreage by $(\hat{\theta}_{ijP}, \hat{\sigma}_{ijP})$, of total production by $(\hat{\theta}_{ijG}, \hat{\sigma}_{ijG})$, and the DE and SE of yield by $(\hat{\theta}_{ijY}, \hat{\sigma}_{ijY})$. Harvested acreage and yield are considered since production could be estimated from these. Data analysis of the CAPS responses for year 2016 resulted in 6 counties with valid DEs for total planted acreage ($\hat{\theta}_{ijP} > 0$) and corresponding SEs equal to zero ($\hat{\sigma}_{ijP} = 0$), 6 counties with valid DEs for total harvested acreage ($\hat{\theta}_{ijH} > 0$) and corresponding SEs equal to zero ($\hat{\sigma}_{ijH} = 0$), and 5 counties with valid estimates DEs for total production ($\hat{\theta}_{ijG} > 0$) and corresponding SEs equal to zero ($\hat{\sigma}_{ijG} = 0$). Direct estimates from data analysis of the CAPS responses for corn yield in 2016 include 104 counties with valid estimates $\hat{\theta}_{ijY} > 0$ and SEs (of the estimates) equal to zero ($\hat{\sigma}_{ijY} = 0$), and 137 counties with both direct estimates positive and SEs below 1 ($\hat{\sigma}_{ijY} < 1$). In total, 241 counties have survey estimated SEs of yield that are smaller than 1, which is nearly 10 percent of all counties. Furthermore, NASS only publishes yield to the nearest 10th of a bushel per acre, and 215 counties have SEs estimated from CAPS smaller than 0.01 bushel per acre. Final model estimates are not produced for the counties with valid (positive) survey estimates and zero/below threshold SEs. The question of interest is: How to mitigate zero/below threshold standard errors? This paper addresses this question through different approaches.

2.2 Subarea Level Model

The subarea level model considered in this paper was first developed by Fuller and Goyeneche (1998) and later studied in a frequentist framework by Torabi and Rao (2014). Erciulescu et al. (2018) study the model in a Bayesian framework, assuming normal distributions in both, the sampling and the linkage models. The authors add a hierarchical level to the Fay and Herriot (1979) area level model by adopting non-informative, proper, independent prior distributions for parameters $\boldsymbol{\beta}', \sigma_u^2, \sigma_v^2$. For completeness, the model from Erciulescu et al. (2018) is described.

The linking model for the true subarea parameters of interest (mean, total, ratio), θ_{ij}, is a linear mixed model:

$$\theta_{ij} = \boldsymbol{x}_{ij}'\beta + v_i + u_{ij} \tag{2.1}$$

The sampling model, for the sample subarea parameters, $\hat{\theta}_{ij}$ is:

$$\hat{\theta}_{ij} = \theta_{ij} + e_{ij}. \tag{2.2}$$

Then, the subarea model for the sample subarea parameters, $\hat{\theta}_{ij}$, is a combination of (2.1) and (2.2):

$$\hat{\theta}_{ij} = \boldsymbol{x}_{ij}'\beta + v_i + u_{ij} + e_{ij}, \tag{2.3}$$

where $i = 1, \dots, m, j = 1, \dots, n_{ci}$. The following assumptions are made:

1. $\hat{\theta}_{ij}|\left(\theta_{ij}, \sigma_{ij}^2\right) \sim N\left(\theta_{ij}, \hat{\sigma}_{ij}^2\right), j = 1, \dots, n_{ci}$
2. $v_i|\sigma_v^2 \sim N(0, \sigma_v^2), i = 1, \dots, m$
3. $u_{ij}|\sigma_u^2 \sim N(0, \sigma_u^2),\ j = 1, \dots, n_{ci}, i = 1, \dots, m$

In our case study, θ_{ij} is the county level parameter of interest. There are m districts (areas) within a state and n_{ci} counties (subareas) within each district i, $\sum_{i=1}^{m} n_{ci} = n_c$. Data consist of survey summaries $\left(\hat{\theta}_{ij}, \hat{\sigma}_{ij}^2\right)$ and auxiliary data x_{ij} for county j of districts i; $\boldsymbol{x}_{ij} = \left(\mathbf{1}, x_{ij}\right)$. This hierarchical Bayesian subarea model is applied to any state for a given year and commodity, and allows for an agreement between the respective estimates at the county level and the district level. Erciulescu et al. (2018) specify priors on the parameters σ_v^2, σ_u^2 and β, and discuss different scenarios of data availability. The joint likelihood and the posterior model are derived under the assumption that both survey and auxiliary data are available. The resulting posterior mean, $\tilde{\theta}_{ij}^{\mathrm{ME}}$, when the direct estimates $(\hat{\theta}_{ij}\hat{\sigma}_{ij}^2)$ and the auxiliary information $\boldsymbol{x}_{ij}$ are both available at the county level is then computed:

$$\tilde{\theta}_{ij}^{\mathrm{ME}} = \tilde{\gamma}_{ij}\hat{\theta}_{ij} + \left(1 - \tilde{\gamma}_{ij}\right)\left\{\boldsymbol{x}_{ij}'\tilde{\beta} + \tilde{\gamma}_i\left(\hat{\theta}_i^{\gamma} - \boldsymbol{x}_i'^{\gamma}\tilde{\beta}\right)\right\} \tag{2.4}$$

where $\tilde{\gamma}_{ij} = \tilde{\sigma}_u^2/\left(\tilde{\sigma}_u^2 + \hat{\sigma}_{ij}^2\right)$, $\tilde{\gamma}_{i.} = \sum_{j=1}^{n_{ci}} \tilde{\gamma}_{ij}$, $\tilde{\gamma}_i = \frac{\tilde{\sigma}_v^2}{\tilde{\sigma}_v^2 + \tilde{\sigma}_u^2 \tilde{\gamma}_{i.}^{-1}}$, $\hat{\theta}_i^{\gamma} = \tilde{\gamma}_{i.}^{-1} \sum_{j=1}^{n_{ci}} (\tilde{\gamma}_{ij}\hat{\theta}_{ij})$, $x_i^{\gamma} = \tilde{\gamma}_{i.}^{-1} \sum_{j=1}^{n_{ci}} (\tilde{\gamma}_{ij}x_{ij})$, and $\boldsymbol{x}_{ij}' = (1, x_{ij})$.

The final modeled county estimates in (2.4) are linear combinations of the DEs and auxiliary data, with coefficients depending on the variances (or SEs) of the DEs. With this approach, counties with valid DEs ($\hat{\theta}_{ij} > 0$) and missing or zero SEs are not modeled. Hence, even though auxiliary data are present, they would not be used and the direct estimate would be reported as the final estimate without a measure of uncertainty. In the rest of the paper, different approaches of mitigating SEs estimated as zero from CAPS are presented. In section 3, the relationship between direct survey estimates for corn in 2016 and their standard errors for all sampled US counties is explored. In the following sections, emphasis is on the SE of the DE of yield.

3. Exploring Relationships between Direct Estimates of Totals

The exploratory data analysis of CAPS survey summaries for corn for sampled US counties revealed a strong relationship (on a log scale) between DEs and their SEs (SEs^2) for total planted acreage, total harvested acreage and production (Figure 1). However, the plot of SEs of yield against the direct estimates of yield for sampled US counties, shown in Figure 1, does not suggest any relationship that could be modeled using classical regression approach.

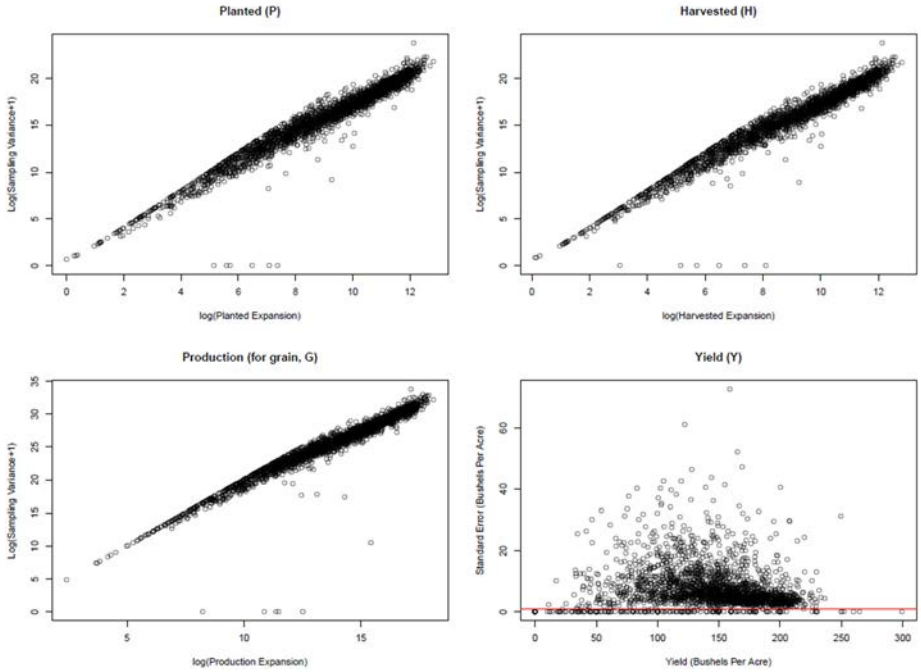

Figure 1: Direct estimates of total planted, harvested, production and yield for sampled US counties in log scale

As an illustration, model (3.1) explains the relationship between the transformed DE of harvested acreage for corn, $\hat{\theta}_{ijH}$ and its transformed variance ($\hat{\sigma}^2_{ijH}$) over all sampled US

counties for year 2016. The multiple R-squared is 0.9666, and the adjusted R-squared is 0.9666.

$$log\left(\hat{\sigma}_{ijH}^2\right) = 1.581039 log\left(\hat{\theta}_{ijH}\right) + 1.532810 \quad (3.1)$$

Model (3.1) was used to approximate (and replace where applicable) survey variances less than the threshold of 1 bu/acre, given that the DEs were positive. A similar relationship was observed between the direct survey estimates for total planted acreage, total harvested acreage and production, and their standard errors. A hierarchical lognormal model for the survey variances on the survey estimates for harvested acreage, corn production and planted area are developed in Erciulescu et al. (2018) to mitigate zero variances. The coefficients are estimated using the subset of sampled data with estimates available (positive) for both quantities. The lognormal model assumption does not hold for yield.

The focus of this paper is on mitigating zero/below threshold SEs (SEs^2) using alternative approaches other than modeling. Plots of several transformations of SEs of the DEs of corn yield did not indicate any relationship that could be modeled well. The distribution of SE of yield has a heavy right tail (Figure 2(a)), so the square root transformation made data more symmetric (Figure 2(b)); however, it did not improve the relationship between SEs and the DEs. The square root transformed SEs of yield are plotted against their corresponding DEs of yield over all sampled US counties in Figure 3. Solid dots represent counties with just 1 or 2 positive reports of yield, and the solid horizontal line denotes 1 bu/acre. Hence, our intervention focused on counties with a small number of reports and/or a small yield. Alternative approaches of mitigating zero estimated SE of yield when the DE is valid (positive) are explored in the next section.

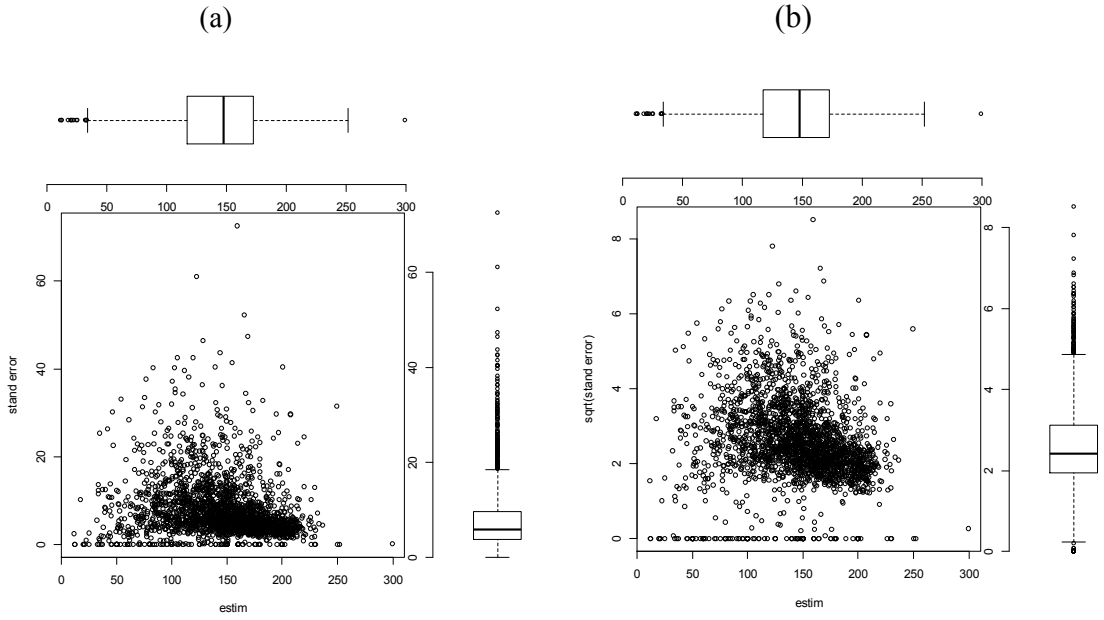

Figure 2: SE (a) and square root transformed SE (b) of yield plotted against their corresponding DE of yield.

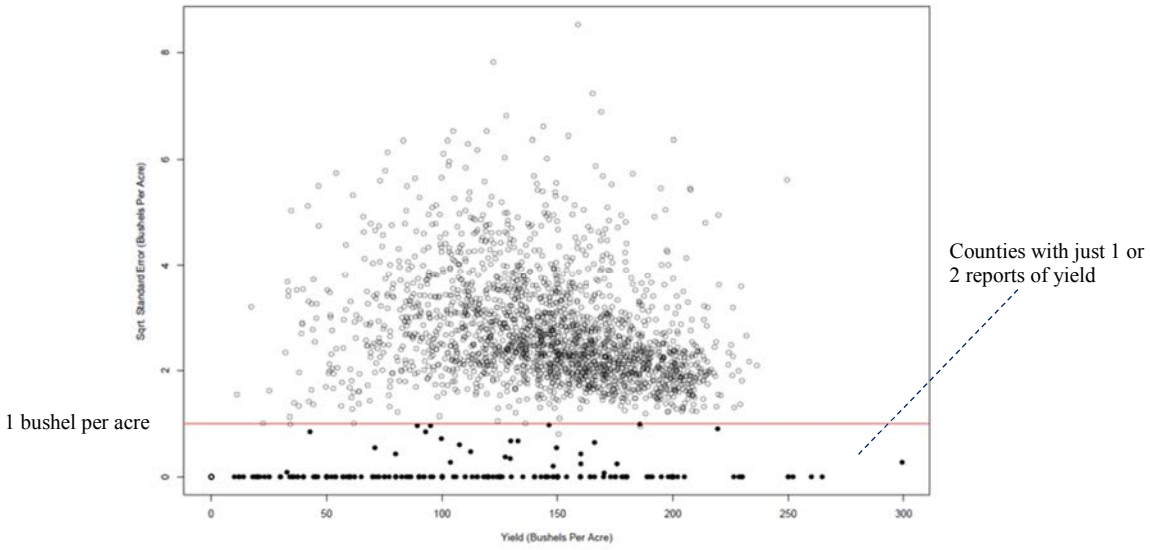

Figure 3: Square root transformed SE of yield plotted against their corresponding DE of yield. Solid line indicates 1 bu/acre and solid dots correspond to counties with just 1 or 2 reports of yield.

4. Alternative Approaches to Standard Errors of Yield

As part of the CAPS summary, variances for DEs of yield based on CAPS are produced using a second order Taylor series approximation, which due to various reasons (e.g., sparseness in data) could result in zero or small estimated variances for several counties. Taylor's approximation is further used with slight modification as an alternative approach to replace the zero/below the threshold variances estimated from CAPS. Other approaches include data driven techniques for estimating the distribution of SEs of yield, using a subset of counties that contain similar information (with respect to the range of covariates already included in the model) as the set of counties with positive DE and SE less than the threshold. We called this an 'enriched' sample data. Illustrated with CAPS data, all approaches are used to approximate (or replace where applicable) survey variances (SE^2) less than 1 bu/acre.

4.1 Taylor's Approximation

A modification of Taylor's approximation is the first alternative approach to replacing the zero/below the threshold variances estimated from CAPS considered. The approximation (4.1) uses the 'imputed' variances for harvested acreage and production based on the lognormal model (3.1) and the correlation between the DEs for harvested acreage and production at the county level approximated by 0.9943, the median of the correlations for all the sampled counties,

$$Var(\hat{\theta}_{ijY}) \approx \left[\frac{E(\hat{\theta}_{ijG})}{E(\hat{\theta}_{ijH})}\right]^2 \left(\frac{Var(\hat{\theta}_{ijG})}{[E(\hat{\theta}_{ijG})]^2} - \frac{2cov(\hat{\theta}_{ijG},\hat{\theta}_{ijH})}{[E(\hat{\theta}_{ijG})][E(\hat{\theta}_{ijH})]} + \frac{Var(\hat{\theta}_{ijH})}{[E(\hat{\theta}_{ijH})]^2}\right). \quad (4.1)$$

For most counties, the $cov(\hat{\theta}_{ijG}, \hat{\theta}_{ijH})$ was computed using the Jackknife method, as part of the survey summary. For the counties where Jackknife method did not produce valid (positive) SEs, $cov(\hat{\theta}_{ijG}, \hat{\theta}_{ijH}) = 0.9943 * \sqrt{Var(\hat{\theta}_{ijG})}\sqrt{Var(\hat{\theta}_{ijH})}$, where $Var(\hat{\theta}_{ijG})$ and $Var(\hat{\theta}_{ijH})$ are estimated from (3.1).

Estimated variances based on Taylor's approximation have high variability throughout the sample due to the variability of the 'imputed' variances for harvested acreage and production from model (3.1) and the approximated correlation between the DEs for harvested acreage and production, propagated through formula (4.1).

4.2 Estimating the Distribution of Standard Errors of Yield

The distribution of SEs of yield is estimated from the 'enriched' sample data. The pool of symptomatic auxiliary variables, already considered as covariates in model (2.3), is chosen by exploring their relationship with the SE of yield. A variable is included in the pool if 1) its correlation coefficient with SE exceeds 0.4 or/and 2) there is a distinct concentration (cluster), in its scatterplot vs SE, of points corresponding to the below threshold standard errors. This pool includes yield DE, production DE, standard error of production and administrative planted acreage values available from the Farm Service Agency (FSA planted acreage) and is used to subset the sample data through two steps as follows:

Step 1: Compute the range of each chosen auxiliary variable, i.e., yield, production, standard error of production and FSA planted acreage corresponding to SEs of yield below 1 bu/acre threshold (illustrated by arrows from A to B in Figure 4).

Step 2: Identify records/units consisting of positive SEs of yield and with values of covariates chosen within the ranges that were computed in step 1 (illustrated by arrows from B to C in Figure 4). This 'enriched' subset of data is used to estimate the distribution of the SEs.

Exploring these data further, it was concluded that the distribution of SEs of yield for in sample US counties can be approximated by a chi-square distribution ($\chi^2_{(a)}$). Under this assumption, the degrees of freedom (a) are estimated using the method of moments. Further, bootstrap sampling combined with a numerical approach are used as another alternative to estimate the degrees of freedom (a), resulting in several densities for approximating the distribution of SEs.

4.2.1 Estimating the parameter of chi-square using method of moments

For the assumed chi-square distribution, using method of moments led to $\chi^2_{(6)}$ as an approximate for the density of SEs of yield (Figure 5). The overlapping densities and the quantile - quantile plot of the original SE of yield and the theoretical $\chi^2_{(6)}$ (Figure 5) show that $\chi^2_{(6)}$ is a good approximation for the SEs of yield for corn estimated from CAPS.

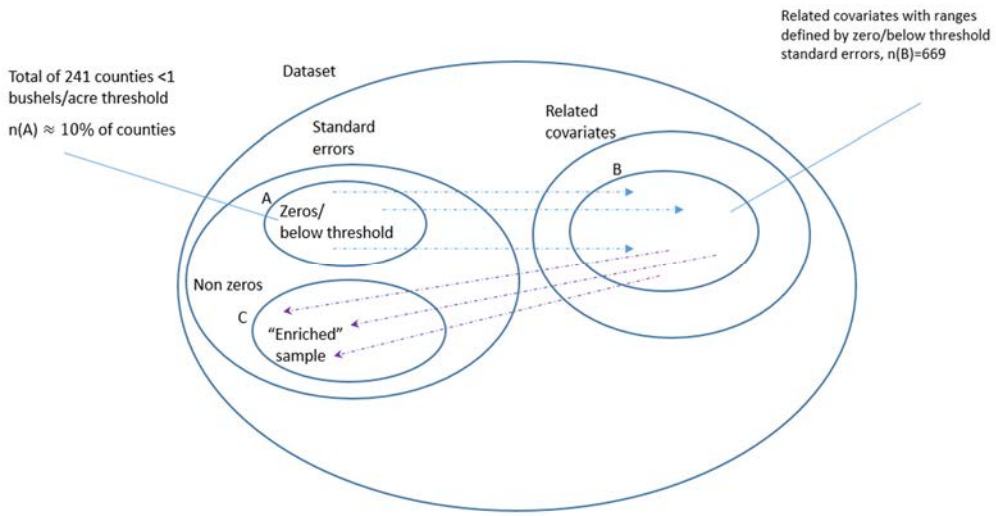

Figure 4: The 'enriched' sample data is defined by zero standard errors of yield and the symptomatic covariates, i.e., yield, production, standard error of production and FSA planted acreage.

4.2.2 Estimating the parameter of chi-square using bootstrap sampling

In this approach, the parameter a is computed numerically based on B bootstrap samples of SEs of yield (from set C in Figure 4). Taking the ratio of the 95% and 5% quantiles of $\chi^2_{(a)}$ and setting it equal to the ratio of the 95% and 5% empirical quantiles estimated from a bootstrap sample of SEs, an implicit equation on a that can be solved numerically is obtained. Repeating the procedure on B bootstrap samples of standard errors, resulted in a sample of size B for a, which is thought of as a realization from the empirical distribution of a. Then, the 75th percentile (a_75) and the max (a_max) of the bootstrap distribution of a are chosen to construct two chi-square distributions, $\chi^2_{(a\_75)}$ and $\chi^2_{(a\_max)}$.

The approximated distributions of SE based on each approach are plotted in Figure 6. A drawing from any approximated distribution of SE based on each approach could be used to replace a below threshold SE of yield estimated from CAPS. The 75th percentile would provide conservatively large SEs as a measure of uncertainty. Also, the empirical 50th and 75th percentiles of SEs within the 'enriched' subset (set C on Figure 4) were used as another empirical approach to approximate the below threshold SE of yield.

In these approaches, a constant (i.e., quantile) from the approximated distribution is used to impute the same value for all counties in the sample, with SEs of yield below the threshold, resulting in a shift in the "pick" of the distribution from zero to around 6 – 10 (Figure 6).

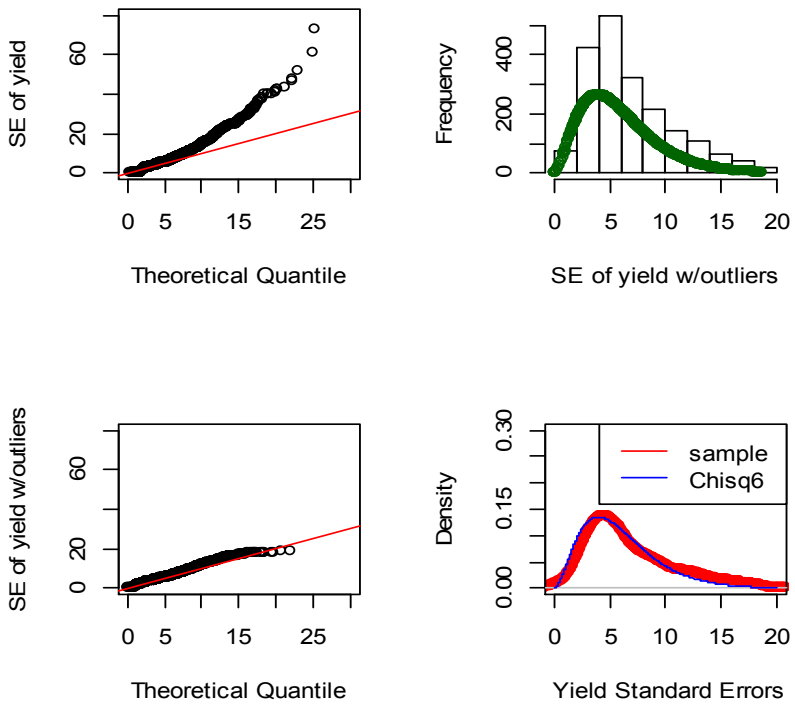

Figure 5: Distribution of positive standard errors in the subspace corn yield in 2016 for the sampled US counties and the density of chi-square(6)

4.3. Comparison of Approaches

Taylor's approximation, drawing of standard errors within an 'enriched' sample data (set C on Figure 4), chi-square approximation of the distribution of SEs, and semi-parametric bootstrap sampling approach provided improvement on the lower tail of the distribution of SE derived from CAPS survey data for corn (Figure 6). All approaches approximate (and replace where applicable) survey variances (SE^2) less than 1 bu/acre, hence providing positive measures of uncertainty for CAPS positive DE. Threshold may differ for other commodities and other years.

The performance of each approach is assessed by measuring the relative "bias" of final model estimates produced using model (2.3), with new imputed variances under each approach described in section 4.2. The difference of final model estimates from the estimates produced by the ASB for selected counties are plotted in Figure 7 with error margins, assuming estimates produced by the ASB as baseline.

Tayler series approximation and the chi-square distribution approach using bootstrap sampling produced final estimates and error margins similar to the ASB estimates for counties with more than one report. For counties with one report and below the threshold positive SE, Tayler series approximation produced higher SEs when compared to the bootstrap approach.

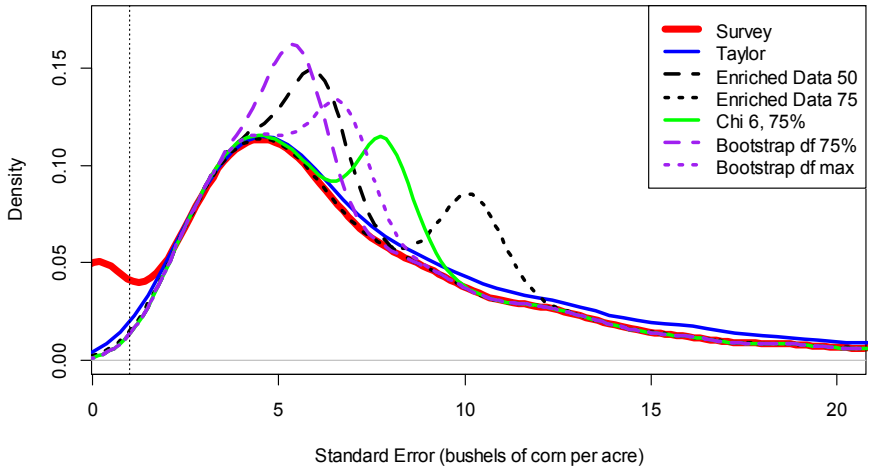

Figure 6: Distribution of standard errors of yield based on different approaches

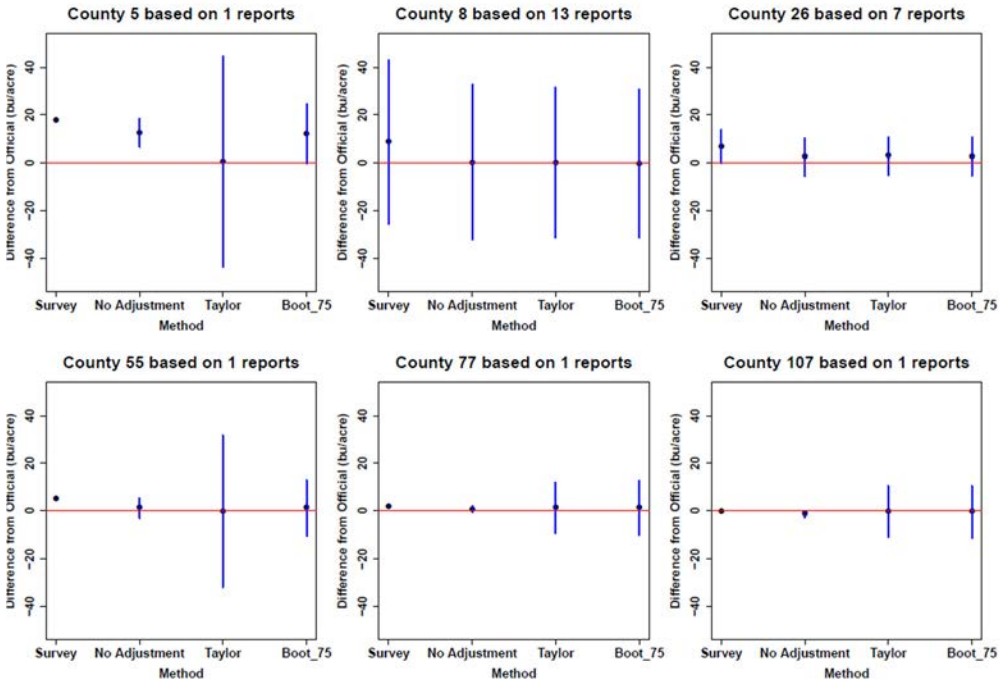

Figure 7: Model predictions of corn yield based on different approaches for counties in selected states

5. Discussion

This paper introduced alternative approaches of mitigating survey variances of yield produced as zeros. NASS's CAPS 2016 data were used to illustrate each approach. The final model estimates for the year 2016 were computed using variances produced from each approach.

The high variability of imputed variances based on Taylor's approximation could be due to the variability associated with this approach using 1) the 'imputed' variances for harvested acreage and production based on a lognormal model (3.1), and 2) the approximated correlation between the DEs for harvested acreage and production (at the county level) i.e., the median of the correlations for all sampled counties. A chi-square with 6 degrees of freedom ($\chi^2_{(6)}$) was a good empirical approximation for the SE of yield for corn. This approach is ad hoc and may be survey specific. The distribution of SE needs to be explored and new distribution assumptions need to be made.

In the bootstrap sampling approach, the degrees of freedom (a) of the assumed chi-square distribution (of SE) were estimated empirically, allowing for some variability on the parameter a. Applying the semi-parametric bootstrap sampling approach to a selected set of counties was as effective as the Taylor's approximation (counties 8, 26, 77 and 107 in Figure 7), and in one case better (counties 55, Figure 7). It would be of interest to explore a fully non-parametric bootstrap approach to estimate the distribution of SE of yield empirically and free of any distribution assumption.

For all approaches, except Taylor's approximation, survey variances less than 1 bu/acre were imputed with a constant, e.g., square of the 75th percentile drawn from the approximated distribution of the SE. Each approach, provided measures of uncertainty for all US counties in the sample with valid (positive) DE of yield. This resulted in an increase by approximately 10% in the number of final model based county estimates with a potential to translate into an increase on the number of counties that NASS publishes official statistics that include measures of uncertainty.

Appendix 1

Estimating the parameter of chi-square using bootstrap sampling

The parameter *a* of $\chi^2_{(a)}$ is estimated numerically based on *B* bootstrap samples of standard errors. The 95% and 5% empirical quantiles estimated from each bootstrap sample are set equal to the ratio of the 95% and 5% theoretical quantiles of the $\chi^2_{(a)}$ distribution.

$$\frac{\chi^2_{(a,.95)}}{\chi^2_{(a,.05)}} = \text{function}(a) = \frac{\hat{x}^{bs}_{.95}}{\hat{x}^{bs}_{.05}}, \qquad (1)$$

where $\hat{x}^{bs}_{.95}$ and $\hat{x}^{bs}_{.05}$ are estimated from the bootstrap sample of SE. Using a normal approximation to $\chi^2_{(a)}$, $\chi^2_{(a)} \approx N(a, 2a)$, the following expression gives an idea about the range of the possible values of a:

$$a = \left\{\left[-\sqrt{2}(z_{.95})\left(1 + \hat{x}^{bs}_{.95}/\hat{x}^{bs}_{.05}\right)\right]/\left(1 - \hat{x}^{bs}_{.95}/\hat{x}^{bs}_{.05}\right)\right\}^2.$$

Solving equation (1) for a numerically for each bootstrap sample of standard errors (of size *m*) results in a sample of length *B* that could be thought of as drawn from the empirical distribution of *a*,

$$a = \text{function}^{-1}\left(\frac{\hat{x}^{bs}_{.95}}{\hat{x}^{bs}_{.05}}\right)$$

Disclaimer

The findings and conclusions in this preliminary publication have not been formally disseminated by the U. S. Department of Agriculture and should not be construed to represent any agency determination or policy.

References

Bell J., and Barboza W. (2012), Evaluation of Using CVs as a Publication Standard. Paper presented at the Fourth International Conference on Establishment Surveys, Montreal, Quebec, Canada, June 11-14.

Cruze N.B., Erciulescu A.L., Nandram B., Barboza W.J., Young L.J. (2016), Developments in Model-Based Estimation of County-Level Agricultural Estimates." ICES V Proceedings. Alexandria, VA: American Statistical Association.

Erciulescu A.L., Cruze N.B., Nandram B. (2016), _Model-Based County-Level Crop Estimates Incorporating Auxiliary Sources of Information." JSM Proceedings. Survey Research Methods Section. Alexandria, VA: American Statistical Association, 3591-3605.

Erciulescu, A.L., and Cruze, N.B., and Nandram, B. (2018). Model-Based

County-Level Crop Estimates Incorporating Auxiliary Sources of Information. *Journal of the Royal Statistical Society*, Series A. https://doi.org/10.1111/rssa.12390.

Ericksen, E. P. (1974). A regression method for estimating populations of local areas. *Journal of the American Statistical Association*, 69, 867-875.

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.

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

Hall, P., & Maiti, T. (2006). On parametric bootstrap methods for small area prediction. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, *68*(2), 221-238. DOI: 10.1111/j.1467-9868.2006.00541.x.

Kott, P. S. (1990). Mathematical formulae for the 1989 survey processing system summary (SPS). https://www.nass.usda.gov/Education_and_Outreach/Reports,_Presentations_and_Conferences/Survey_Reports/Mathematical%20Formulae%20for%20the%201989%20Survey%20Processing%20System%20(SPS)%20Summary.pdf.

Nicholls, A. (1977). A regression approach to small area estimation. Australian Bureau of Statistics. (Mimeographed).

Purcell, N. J. and Kish, L. (1979). Estimation for small domains. *Biometrics*, 35, 365-384.

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

Torabi M. and Rao J.N.K. (2014). On small area estimation under a sub-area level model. *Journal of Multivariate Analysis*, 127, 36-55.

Tzavidis N., Zhang, L-C., Luna Hernandez, A., Schmid, T., and Rojas-Perilla, N. (2018). From start to finish: a framework for the production of small area official statistics. *Journal of the Royal Statistical Society*, Series A, 181, Part 4, 1-33.

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