Toward an updated publication standard for official county-level crop estimates

Nathan B. Cruze*, Andreea L. Erciulescu†*, Habtamu K. Benecha*,
Valbona Bejleri*, Balgobin Nandram‡*, Linda J. Young*

Abstract

Statistical agencies face demands for official estimates at increasingly finer levels of disaggregation. The United States Department of Agriculture's National Agricultural Statistics Service (NASS) is striving to meet such demands. NASS first implemented nationwide probability sampling for its crops county estimates in 2011. The underlying probability sample for crops county estimates has been an integral part of the production of its official statistics. The existing publication standard is based on either having sufficient number of reports (30 or more responses for that crop) or sufficient coverage (the sum of unweighted reports accounts at least 25% of the estimated harvested area total). NASS is transitioning to the use of model-based county estimates. For publication, these estimates need to protect confidentiality and be precise enough to be fit for use. At the same time, the publication standard must be easily described to NASS's clientele, who include farmers, researchers, other USDA agencies, and Congress. In this paper, we discuss the challenges and intricacies of developing such a publication standard.

Key Words: Disclosure; fit for use; publication standard; small area estimation; synthetic

1. Introduction

Like statistical agencies worldwide, the United States Department of Agriculture's (USDA's) National Agricultural Statistics Service (NASS) faces increasing demand for statistics published at finer levels of detail. Iwig (1996) notes that USDA's first involvement in publishing county-level agricultural estimates dates back to 1917, when the department, through federal-state cooperative agreement, assisted the Wisconsin State Board of Agriculture with the preparation of county-level crop estimates. Additional state cooperative agreements were authorized soon thereafter, reflecting the value and importance that local agricultural estimates provide to the agricultural sector and the public in general.

More than one century later, NASS continues annual publication of official crop county estimates for dozens of federally-mandated and state-funded small grains and row crops. These statistics are published at two smaller-than-state domains: the agricultural statistics district, and the county. Agricultural statistics districts are comprised of one or contiguous counties within the same state. Where warranted, finer detail may still be desired. For select states and commodities, NASS publishes sub-domain estimates by practice (e.g., irrigated versus non-irrigated practice), purpose (e.g., for grain and for silage), or specific types of the same commodity crop (e.g., pinto, garbonzo, and other types of beans).

In addition to NASS's own survey data, a variety of auxiliary data are available through administrative sources provided by other USDA agencies. Remote sensing technologies such as NASS's Cropland Data Layer (CDL) discussed in Boryan et al. (2011) represent additional sources of data still, and new data types such as precision agricultural data are on the horizon. NASS's traditional process for synthesizing various sources of information

*USDA National Agricultural Statistics Service, 1400 Independence Ave., SW, Washington, DC 20250

†National Institute of Statistical Sciences, 1750 K Street, NW, Suite 1100, Washington DC 20006-2306

‡Worcester Polytechnic Institute, Department of Mathematical Sciences, Stratton Hall, 100 Institute Road, Worcester, MA 01609

has been through the expert opinion of its Agricultural Statistics Board (ASB). The role of the ASB in several contexts is described in Allen (1992); precursors to NASS's current procedures for producing county estimates are discussed in Iwig (1996).

In an effort to continue to improve its processes and products, NASS engaged the National Academies of Sciences, Engineering, and Medicine Committee on National Statistics for an external review of its current practices. In late 2017, the external panel completed its charge with the publication of a consensus study report titled *Improving Crop Estimates by Integrating Multiple Data Sources* (National Academies of Sciences, Engineering, and Medicine, 2017). As the panel learned, NASS county estimates are used for many purposes including county-loan administration, crop insurance, and other agricultural safety net programs. The recommendations of the panel are made with a view toward taking advantage of multiple sources of data in a transparent, structured manner, while increasing the number of fit-for-use estimates for a variety of purposes. Some of these recommendations entail changes to the way NASS has typically produced county-level crop estimates and the standard used to evaluate them for publication.

In Section 2, we summarize some of the current practices, publication standard, and confidentiality criteria used at NASS. The findings of the external review are likely to change the way NASS assimilates multiple sources of data in the future. In Section 3, we discuss and interpret several of the panel's recommendations as they pertain to a publication standard for modeled smaller-than-state crop estimates of acreage, production, and yield. Concluding remarks are offered in Section 4.

2. Current Practices

2.1 Current Procedures for Combining Multiple Sources of Data

The ASB process can be described as 'top-down', both in terms of the sequence of publication of national, state, and sub-state crop estimates, and in the manner that estimates for agricultural statistics districts and county estimates are constructed and reconciled. Quarterly Acreage, Production, and Stocks (APS) surveys are conducted in support of end-of-season national and state estimates. The September and December APS surveys capture the seasonality and completion of harvest of major small grains and row crops, respectively. Supplemental, post-harvest samples collected under the County Agricultural Production Survey (CAPS) augment the list-based sample obtained from APS; essentially the APS survey and CAPS act as a single, reweighted sample obtained under a single, multivariate probability proportional to size (MPPS) survey design (Bailey and Kott, 1997). Hereafter, we use the acronym CAPS to refer to direct estimates constructed using the pooled, list-based data collected from both surveys.

Data collection for CAPS is ongoing, even as NASS publishes official state estimates for planted area, harvested area, production, and yield. Thus, official external targets for state-level totals are given at the time that official district and county estimates are to be set. Estimates are established state-by-state for each type of sampled crop. Beginning with the agricultural statistics districts within a given state, a NASS field office staff member reviews separate CAPS survey and non-survey (administrative or remote sensing) inputs and sets estimates on the three totals in the following order: planted area, harvested area, and production. This reflects a 'conditional' logic; given that planted area exists, how much area was successfully harvested, and given that acreage was harvested, what was the total output? The panel learned that staff members use informal composites to combine multiple sources of data. The composite, district-level estimates are then benchmarked to the state totals and rounded in accordance with rounding rules. Subsequently, the process is repeated

at the county level, and benchmarked, rounded county estimates of planted area, harvested area, and production are produced. Official yield estimates are obtained as the ratio of final production estimates to final harvested area at both levels. The ASB facilitates a review of these estimates, ensuring that proper procedures were followed, and that the NASS publication standard is enforced. More detail about the ASB procedures, available inputs, and composite weighting can be found in (National Academies of Sciences, Engineering, and Medicine, 2017, pp. 23-26).

2.2 Current Publication Standards

Due to changes in year-to-year planting decisions, not every sampled unit necessarily has the crop of interest. Direct estimates for a county (or district) are constructed from the number of positive reports, i.e., reports from respondents who affirm they have non-zero amounts of the targeted crop(s). Item-level nonresponse in the survey data is possible; a respondent could provide acreage data, but decline to provide data on the associated yield or production. It is also possible for a farmer, to plant the crop of interest but be unable to harvest any portion of it due to severe drought or other reasons, in which case his data would be a valid zero. Thus, the number of positive reports used to estimate yield or production, denoted n^+_{yield}, may be smaller than the number of positive reports for planted area, $n^+_{planted}$. (NASS accounts for this through reweighted estimators.) This possibly smaller number of yield reports n^+_{yield} figures into NASS's publication and disclosure criteria in several ways. NASS evaluates estimates for publication according to two criteria verified in the following order:

1. minimum number of reports: $n^+_{yield} \geq 30$ **or**
2. for counties with $n^+_{yield} < 30$, harvested area coverage $\geq 25\%$, where

$$coverage \equiv \frac{\sum_{i \in n^+_{yield}} (unweighted\ havested\ report)_i}{ASB\ harvested\ area\ estimate}. \tag{1}$$

This standard is applied to all smaller-than-state estimates, including agricultural statistics districts and any other aggregates of counties that may result.

In addition to its publication standard, NASS has a number of confidentiality procedures in place. In particular NASS suppresses estimates according to the following criteria:

- there are $n^+_{yield} < 3$ reports in the county,
- a 'dominant' operation accounts for a large proportion of a county's estimated total acreage or production ($\geq p\%$, not to be disclosed),
- or a county's total planted acreages fall below a specified threshold that may vary by commodity.

Since NASS publishes nested totals (county within district within state), complementary suppression is practiced to prevent disclosure of suppressed estimates that could be learned by simple subtraction. By similar reasoning, acreage, production, and yield estimates are related quantities, i.e., planted area$\geq$ harvested area, and yield=production$\div$harvested area; NASS publishes all of these estimates for a county, or it publishes none to prevent disclosure of estimates by simple arithmetical operations. Thus, counties that may otherwise be suitable for publication are suppressed in the interest of maintaining confidentiality.

The publication standards noted above were conceived in 2008, prior to the pilot activities that gave rise to the CAPS program. In 2009 and 2010, four pilot states were sampled

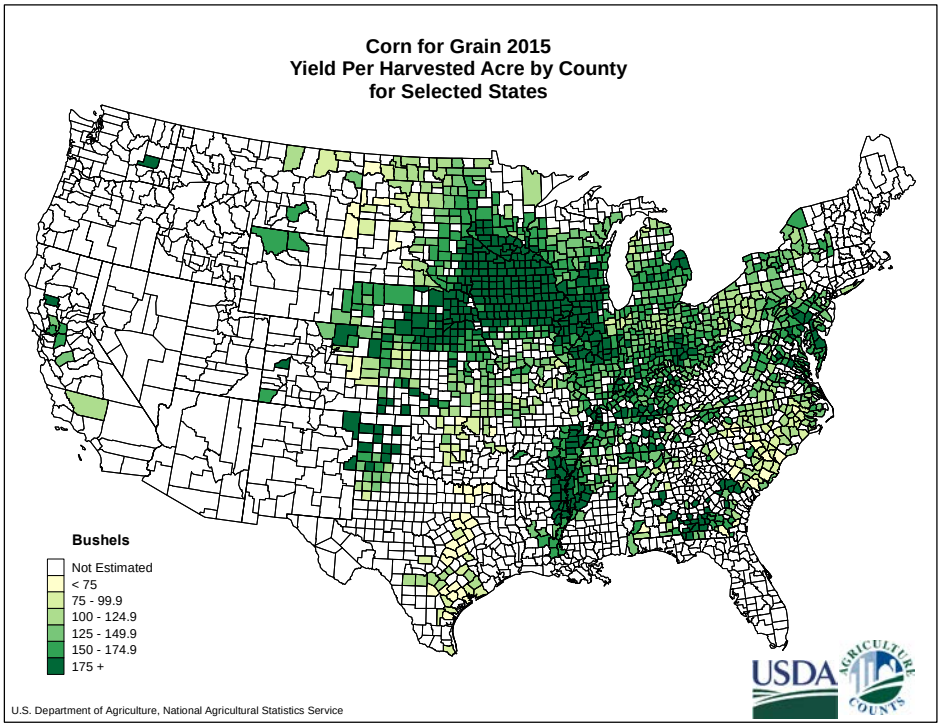
under a multivariate probability proportional to size (MPPS) design in support of county estimates (Bailey and Kott, 1997). In 2011, NASS successfully implemented nationwide probability sampling. Prior to the inception of CAPS, the county estimates program essentially operated as "45 different programs conducted separately by each NASS State Statistical Office (SSO)." (Iwig, 1996, Ch. 7, p. 3). While a basic set of operating principles was provided to each SSO, (including the aforementioned disclosure criteria regarding the minimum number of reports and dominant operation thresholds), state offices exercised some latitude in the upkeep of list frames, the collection and analysis of (non-probability) survey data, and the evaluation of the fitness of resulting estimates for purpose and publication. Along with a standardized survey design, the transition to a probability-based sample represented a commitment to a more standardized list frame upkeep and data collection procedures across the nation; the adoption of the publication standard represented a commitment to a more uniform evaluation of an estimates suitability for publication.

NASS's current publication standard was developed when parts of the county estimates program were to be collected under a combination of probability and non-probability surveys. For select crops, Table 1 bins counties by number of corresponding positive reports obtained during the 2015 crop year. The number of counties with at least one planted report is often smaller than the total number of counties within program states, reflecting the fact that a crop may not be present in every county in a state. For example, 36 states comprised of 2,837 counties were sampled is support of corn for grain county estimates. The presence of corn was affirmed by NASS surveys with positive planted data in 2,426 counties. Of those, 1,948 $(1,254 + 694)$ had at least three associated yield reports. Under the current publication standard, only 1,661 of these counties were candidates for publication, as some counties failed the coverage standard. Complementary suppression and other confidentiality concerns reduced the number of individually published corn for grain counties to 1,433. Given NASS's current stance on disclosure, a considerable number of county estimates were suppressed on the basis of fewer than three positive yield reports; some 50% of sorghum estimates were suppressed on that basis, irrespective of other qualities of the estimates.

Table 1: Binning counties by number of positive reports, 2015 crop year

Crop	Corn in 36 states	Soybeans in 30 states	Winter Wheat in 33 states	Sorghum in 12 states
Total counties	2,837	2,563	2,597	1,130
Counties with $n^+_{planted} > 0$	2,426	2,012	2,191	754
Counties with $0 \leq n^+_{yield} < 3$	478	332	696	372
Counties with $3 \leq n^+_{yield} < 30$	1,254	928	1,245	335
Counties with $n^+_{yield} \geq 30$	694	752	250	47
Published counties	1,433	1,306	1,048	218

The maps shown in Figure 1 contrast the number of published county estimates in the official product (top panel) with the number of confided estimates (bottom). The current NASS publication standard does give counties with a small number of reports (3 to 29) the possibility of being published, however, if this standard is unduly stringent, the result could be a high rate of suppression. While the suppressed counties may be represented in combined county or agricultural statistics district aggregates, these estimates may lack the specificity or detail desired for the administration of local agricultural policy or other purposes.



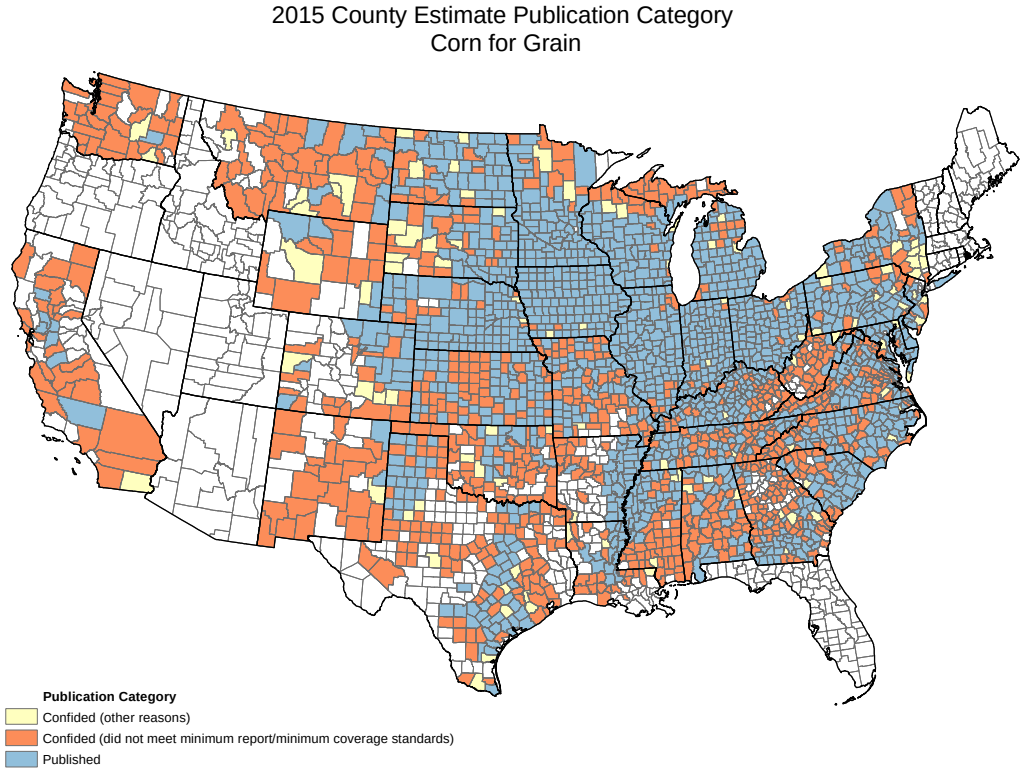


Figure 1: Top: official choropleth map of corn for grain yield. Such maps are released annually at `https://www.nass.usda.gov/Charts_and_Maps/Crops_County/index.php`. Bottom: additional counties suppressed in accordance with NASS publication standards.

3. Recommendations of the External Review

Improving Crop Estimates by Integrating Multiple Data Sources (National Academies of Sciences, Engineering, and Medicine, 2017) presented a vision for incorporating a variety of current and emerging data sources with a goal of improving the transparency and reproducibility of NASS business processes and becoming more responsive to data users' needs. The proposed changes in key recommendations include evolving the role of the ASB (Recommendation 2-1), utilizing well-documented models to synthesize multiple data sources (2-2), and publishing measures of uncertainty for all estimates (2-4). One recommendation in particular entails revising the publication standard to reflect uncertainty in the estimates:

> Recommendation 2-3: The National Agricultural Statistics Service (NASS) should adopt the following publication standard:
>
> - County-level estimates may be withheld to protect confidentiality.
> - County-level estimates may be withheld because NASS deems them unreliable for any use, based on its measure of uncertainty.
> - All other county-level estimates will be published, along with their measures of uncertainty. (National Academies of Sciences, Engineering, and Medicine, 2017, p.3).

Collectively, all four recommendations (2-1, 2-2, 2-3, 2-4) presuppose that a suitable small area methodology or other model-based approach is already available. Recommendation 2-3 leaves room for interpretation. Which measures of uncertainty should NASS publish? Should the same thresholds be applied to both totals and the yield ratio? Should the standard be tied to a key estimate? Can a model that is composed of both survey and 'other' data sufficiently protect counties with a small number of respondents or with a dominant operation. Internally at NASS, the meanings of two key phrases found in the panel's report are being discussed: *fit for use* and *suitably synthetic*. Using corn for grain data collected in during the 2015 crop year, we apply a variation of the Bayesian, subarea model methodology developed in Erciulescu et al. (2018b); the results serve to explore notions of fitness for use and disclosure avoidance as they relate to any candidate modeled estimate.

3.1 Fit for Use

The panel's recommendations suggest that fitness for use is tied to the uncertainty of estimates. Within one year of implementing CAPS nationwide, Bell and Barboza (2012) explored the relationship of coefficients of variation of *direct estimates* in relation to the NASS publication standard, noting that in some cases, the coverage threshold permitted publication of counties with direct estimates subject to high relative variability. One caveat in that analysis is that the NASS official statistic may be an implicit *shrinkage estimate*. Through the ASB process, auxiliary information is brought to bear, although its value is not quantified in a measure of uncertainty.

In Figure 2, coefficients of variation for harvested area and yield estimates are plotted against number of reports used to construct the direct estimates. In the left panel, points corresponding to survey harvested area totals (black) show somewhat higher coefficients of variation than their counterparts produced under the model (red); a well-defined methodology can give rise to measures of uncertainty, capture the contribution of administrative (or other auxiliary) data, and result in a larger number of more precise estimates. In the right panel, yield estimates under survey and model show more comparable performance. Since yield is the ratio of production to harvested area, and those totals are highly positively

correlated, the survey estimates for yield already tend to have lower relative variability. Nevertheless, models that borrow strength from other counties and districts within the state and incorporate a crop productivity index can bring about some reduction in the uncertainty associated with the county yield estimates.

The pattern shown in the harvested area panel of Figure 2 may be indicative of the planted area and production totals as well. Three popular thresholds (20%, 30% and 50%) are plotted. A comparison of the harvested area totals and the yield estimates in terms of their relative variability raises another important point. Insisting that all estimates (total or ratio) conform to a common threshold would result in a smaller number of estimates, likely determined by estimated totals. Some agencies distinguish between key estimates and secondary estimates, verifying publication standards for the key estimates, and allowing secondary estimates to be published even if the same criteria are not met. (U.S. Census Bureau, 2013, Requirement F-1, p. 114) Declaring the yield estimate a 'key estimate' would likely result in a higher number of published estimates, however some associated estimated totals in the county could be subject to high relative variability. Other strategies still could tie the standard to average or median measures of uncertainty taken across states or districts, or publishing all estimates with measures of uncertainty and letting end users determine their fitness for use. Given that NASS official estimates are often used to administer payments made by other USDA agencies under some farm programs, this is an important determination still to be made in communication with NASS's many stakeholders.

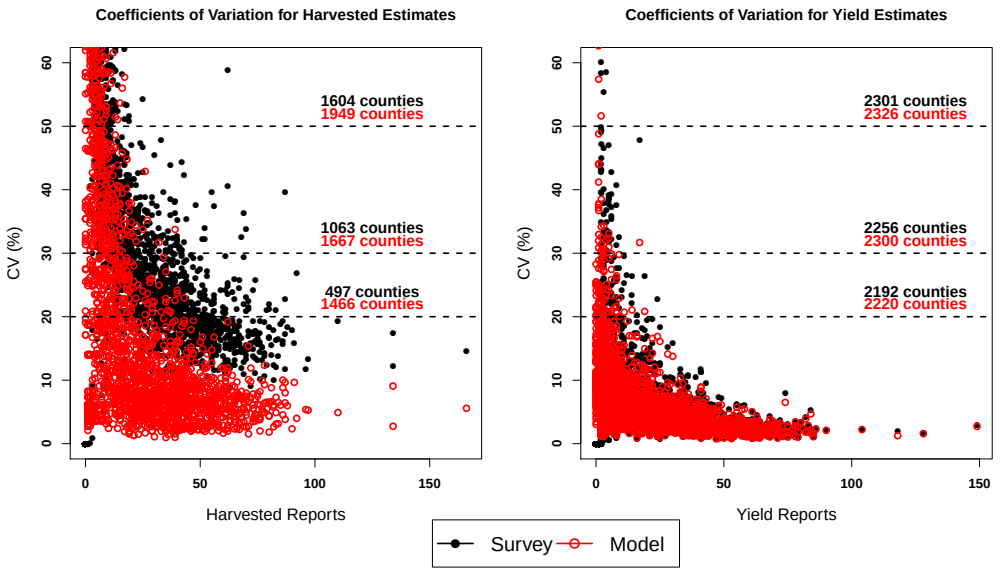


Figure 2: Comparison of coefficients of variation under survey and model for the harvested area totals and yield ratios (corn for grain, 2015)

3.2 Suitably Synthetic

Like other statistical agencies in the U.S. and around the world, NASS has statutory obligations to protect the data of individual respondents. Assurances legislated under the Confidential Information Protection and Statistical Efficiency Act of 2002 (CIPSEA) apply to federal agencies more broadly; Title 7 of the U.S. Code pertains to the Department of Agriculture in particular. Both of these statutes preclude the release of *individually identifiable* data without consent.

On the matter of estimates incorporating survey and non-survey data, the panel opined:

> If the farm provided data under pledge of confidentiality, could [NASS] then publish an estimate derived entirely from other sources? Officially the answer would be yes, but how would the farmer view this publication?
> (National Academies of Sciences, Engineering, and Medicine, 2017, p.39)

In this statement, the distinction between disclosure and inferential or perceived disclosure risk can be seen. The latter question matters because respondents participate in CAPS and many other NASS surveys on a voluntary basis.

Lokupitiya et al. (2007) describe a methodology for reconstructing a 16 year history of county estimates for crop acreage and yield, including estimates for unpublished counties. They used regression analyses and linear mixed-effect models to combine historical NASS published estimates, historical Census of Agriculture, and a number of local environmental factors. While they cannot claim to have recovered record-level data, they do claim to achieve $\pm 10\%$ absolute relative error when compared to NASS official estimates. In areas with a particularly small number of survey respondents or a dominant operation, this might represent some risk of inferential disclosure.

The experiences of other statistical agencies in the United States may be informative. The Census Bureau's Small Area Income and Poverty Estimates are a series of completely model-based official estimates indicating counts and proportions of children in poverty in school districts and counties (U.S. Census Bureau, 2017) . The panel notes that, to date, the Census Bureau's Disclosure Review Board has never opted to suppress these small area estimates, using the phrase 'suitably synthetic' to describe estimates that protect sensitive American Community Survey data (*a household survey*), aggregated tax records, and other administrative data. In the opinion of the Disclosure Review Board, "given that [these data are] fully synthetic, the synthesis itself is the disclosure avoidance protection technique." (National Academies of Sciences, Engineering, and Medicine, 2017, p. 40)

In a companion paper appearing in this proceedings, Erciulescu et al. (2018a) show that the estimated posterior mean can be expressed as a weighted average of survey and 'other' data, a synthesis that may help avoid disclosure of respondent data. Letting $\tilde{\theta}_{ij}$ denote the estimated posterior mean for the j^{th} county estimate within the i^{th} agricultural statistics district, the estimate decomposes as a weighted average of the county's direct estimate $\hat{\theta}_{ij}$ and auxiliary information contained in the bracket term,

$$\tilde{\theta}_{ij} = \tilde{\gamma}_{ij}\hat{\theta}_{ij} + (1 - \tilde{\gamma}_{ij}) \left\{ \boldsymbol{x}'_{ij}\tilde{\boldsymbol{\beta}} + T \right\}. \tag{2}$$

In Equation 2, the term T represents a more complex expression for information borrowed from other counties within the same district, a potential benefit of the subarea-level model. (In a Fay-Herriot model, $T = 0$.) Starting with the direct estimate itself may afford some protection as it is a weighted sum (or function of weighted sums) of record-level survey data. The combination with other auxiliary data and information from other counties within the same district might further help protect respondent confidentiality. In Erciulescu et al. (2018a), strategies for using the bracketed term to produce out-of-sample predictions are

investigated to combat the lack of survey data in some counties. Such an estimate is truly synthetic as there may be no survey data in the given county, i.e., no $\hat{\theta}_{ij}$ exists, however administrative data may indicate that producing an estimate for the county is still appropriate. The statutory obligations to protect data likely having been fulfilled, NASS will have to determine as an agency how respondents might view the publication of such estimates, and whether the strategy offers a suitably synthetic protection for any dominant producers within a county in the context of NASS's *establishment survey*.

4. Conclusions

As an agency, NASS is assessing the feasibility of recommendations made in *Improving Crop Estimates* (National Academies of Sciences, Engineering, and Medicine, 2017) and transitioning to a system of model-based crop county estimates. The transition to a system of modeled county estimates is intended to help NASS respond to the needs of its customers, incorporate existing and emerging data sources, and publish measures of uncertainty. As a first step in this transition, agency review of historical crop estimates produced under the candidate methodology of Erciulescu et al. (2018b) is underway. Pending satisfactory model validation and review of model-based estimates reflecting the breadth of commodities surveyed under the NASS crop county estimates program, the agency will have to determine the publication standard that will govern a program of model-based estimates, and communicate those decisions clearly to a wide variety of stakeholders.

Acknowledgments

The authors gratefully acknowledge James Johanson and Ray Roberts of USDA NASS for their assistance in preparing the suppression map in Figure 1. This research was supported in part by the intramural research program of the U.S. Department of Agriculture, National Agricultural Statistics Service.

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 formal determination on policy.

References

Allen, R. (1992). Statistical Defensibility as Used by U.S. Department of Agriculture, National Agricultural Statistics Service. *Journal of Official Statistics*, 8(4):481–498.

Bailey, J. and Kott, P. (1997). An Application of Multiple List Frame Sampling for Multi-Purpose Surveys. In *JSM Proceedings, Survey Research Methods Section*, pages 496–500. American Statistical Association, Alexandria, VA.

Bell, J. and Barboza, W. (2012). Evaluation of Using CVs as a Publication Standard. In *Proceedings of the Fourth International Congress on Establishment Surveys*. American Statistical Association, Montreal.

Boryan, C., Yang, Z., Mueller, R., and Craig, M. (2011). Monitoring U.S. agriculture: the U.S. Department of Agriculture, National Agricultural Statistics Service, Cropland Data Layer Program. *Geocarto International*, 26(5):341–358.

Erciulescu, A., Cruze, N., and Nandram, B. (2018a). Combining Survey and Administrative Data to Produce Official Statistics. In *JSM Proceedings, Survey Research Methods Section*. American Statistical Association, Alexandria, VA. Appearing in this proceedings.

Erciulescu, A., Cruze, N., and Nandram, B. (2018b). Model-Based County-Level Crop Estimates Incorporating Auxiliary Sources of Information. *Journal of the Royal Statistical Society, Series A*. In press, https://doi.org/10.1111/rssa.12390.

Iwig, W. (1996). The National Agricultural Statistics Service County Estimates Program. In Schaible, W., editor, *Indirect Estimators in U.S. Federal Programs*, chapter 7, pages 129–144. Springer, New York.

Lokupitiya, E., Breidt, F., Lokupitiya, R., Williams, S., and Paustian, K. (2007). Deriving Comprehensive County-Level Crop Yield and Area Data for U.S. Cropland. *Agronomy Journal*, 99:673–681.

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

U.S. Census Bureau (2013). *U.S. Census Bureau Statistical Quality Standards*. Washington, D.C. `https://www.census.gov/content/dam/Census/about/about-the-bureau/policies_and_notices/quality/statistical-quality-standards/Quality_Standards.pdf`.

U.S. Census Bureau (2017). Small Area Income and Poverty Estimates (SAIPE) Program: About. `https://www.census.gov/programs-surveys/saipe/about.html`. Accessed: 2018-08-27.