6-2018

Combining Survey and Non-survey Data for Improved Sub-area Prediction Using a Multi-level Model

Jae Kwang Kim
Iowa State University, jkim@iastate.edu

Zhonglei Wang
Iowa State University, wangzl@iastate.edu

Zhengyuan Zhu
Iowa State University, zhuz@iastate.edu

Nathan B. Cruze
U.S. Department of Agriculture

Combining Survey and Non-survey Data for Improved Sub-area Prediction Using a Multi-level Model

Abstract
Combining information from different sources is an important practical problem in survey sampling. Using a hierarchical area-level model, we establish a framework to integrate auxiliary information to improve state-level area estimates. The best predictors are obtained by the conditional expectations of latent variables given observations, and an estimate of the mean squared prediction error is discussed. Sponsored by the National Agricultural Statistics Service of the US Department of Agriculture, the proposed model is applied to the planted crop acreage estimation problem by combining information from three sources, including the June Area Survey obtained by a probability-based sampling of lands, administrative data about the planted acreage and the cropland data layer, which is a commodity-specific classification product derived from remote sensing data. The proposed model combines the available information at a sub-state level called the agricultural statistics district and aggregates to improve state-level estimates of planted acreages for different crops. Supplementary materials accompanying this paper appear on-line.

Keywords
Agricultural survey, Hierarchical model, Mean squared prediction error, Small area estimation, Survey integration

Comments
This article is published as Kim, Jae Kwang, Zhonglei Wang, Zhengyuan Zhu, and Nathan B. Cruze. "Combining Survey and Non-survey Data for Improved Sub-area Prediction Using a Multi-level Model." *Journal of Agricultural, Biological and Environmental Statistics* 23, no. 2 (2018): 175-189. doi: 10.1007/s13253-018-0320-2.

Supplementary materials for this article are available at https://doi.org/10.1007/s13253-018-0320-2.

Combining Survey and Non-survey Data for Improved Sub-area Prediction Using a Multi-level Model

Jae Kwang KIM, Zhonglei WANG, Zhengyuan ZHU, and Nathan B. CRUZE

Combining information from different sources is an important practical problem in survey sampling. Using a hierarchical area-level model, we establish a framework to integrate auxiliary information to improve state-level area estimates. The best predictors are obtained by the conditional expectations of latent variables given observations, and an estimate of the mean squared prediction error is discussed. Sponsored by the National Agricultural Statistics Service of the US Department of Agriculture, the proposed model is applied to the planted crop acreage estimation problem by combining information from three sources, including the June Area Survey obtained by a probability-based sampling of lands, administrative data about the planted acreage and the cropland data layer, which is a commodity-specific classification product derived from remote sensing data. The proposed model combines the available information at a sub-state level called the agricultural statistics district and aggregates to improve state-level estimates of planted acreages for different crops.

Supplementary materials accompanying this paper appear on-line.

Key Words: Agricultural survey; Hierarchical model; Mean squared prediction error; Small area estimation; Survey integration.

1. INTRODUCTION

Combining information from several sources to improve estimates for population parameters is an important practical problem in survey sampling. In the past decade, more and more auxiliary information becomes available, including large administrative record datasets and remote sensing data derived from satellite images. How to combine such information with survey data to provide better estimates for population parameters is a new challenge that survey statisticians face today. Tam and Clarke (2015) present an overview of some initiatives of big data applications in official statistics of the Australian Bureau of Statistics.

Jae Kwang Kim (✉), Zhonglei Wang and Zhengyuan Zhu, Department of Statistics, Iowa State University, Ames, IA 50011, USA (E-mail: *jkim@iastate.edu*). Nathan B. Cruze, National Agricultural Statistics Service, United States Department of Agriculture, Washington, DC 20250, USA.

Journal of Agricultural, Biological, and Environmental Statistics, Volume 23, Number 2, Pages 175–189
https://doi.org/10.1007/s13253-018-0320-2

We first provide a brief review of classical methods in survey statistics for combining information from multiple sources. One way is through calibration weighting, or benchmarking weighting. Perhaps, the first application using calibration weighting was discussed by Deming and Stephan (1940), where the census information is incorporated by the raking ratio method. Zieschang (1990), Renssen and Nieuwenbroek (1997), Dever and Valliant (2010) among others used weighting methods to combine information from different surveys; see Kim and Park (2010) and Wu and Lu (2016) for a comprehensive overview of the calibration estimation in survey sampling. Survey integration is another approach to combine information of several surveys from the same target population. Hidiroglou (2001) proposed to combine two surveys in the context of a non-nested double sampling. Merkouris (2004, 2010) gave a rigorous treatment of the survey integration through the generalized method of moments. Legg and Fuller (2009) and Kim and Rao (2012) developed synthetic imputation approaches to combining two surveys.

Small area estimation is an emerging research area in survey sampling, and it addresses the best prediction problem for small areas where auxiliary information is available outside the sample; see Fay and Herriot (1979), Pfeffermann (2002) and Rao and Molina (2015) for details. A multi-level model is useful for analyzing data with hierarchical structures. Torabi and Rao (2008) proposed a two-level model approach by assuming an equal slope for the entire population. Torabi and Rao (2014) investigated the small area estimation problem by a sub-area-level model based on a linear regression with normal random effects. Datta and Ghosh (1991) used a hierarchical Bayesian approach in a mixed linear model. Raghunathan et al. (2007) discussed a hierarchical Bayesian model by an arcsine transformation for the bounded variable. Datta and Ghosh (2012) proposed a Bayesian shrinkage method to borrow strength for the small area estimation. Manzi et al. (2011) considered a Bayesian model to incorporate the uncertainties of estimates from different surveys. Elliott and Davis (2005) discussed a Bayesian approach to adjust the survey weight using a propensity score method. For more Bayesian methods handling small area estimation problems, see Ghosh and Rao (1994), Datta (2009) and Ghosh et al. (1998).

In this paper, we generalize the model considered by Torabi and Rao (2014) and consider a more comprehensive application for the crop acreage estimation problem. To achieve this goal, we first build a level-one model for each area using area-specific parameters, which account for the heterogeneity among areas. Then, a level-two model for area parameters is developed to borrow strength across areas. Sampling error models are also set up for both levels to link the observed data with latent variables, such as the regression coefficients involved in the level-one and level-two models. A novel frequentist approach is developed to estimate parameters, and it can be generalized to multi-level models. An estimate of the mean squared prediction error is derived through a Taylor linearization. The proposed method is similar in spirit to a hierarchical Bayesian approach applied to the small area estimation problem, but it uses a frequentist approach to estimate model parameters and quantify their uncertainties. A simulation study is conducted to test the proposed method, and it shows that estimates from the two-level model improve the direct survey ones by incorporating auxiliary information. The two-level model is applied to estimate the state-level planted acreage of corn, soybean, and winter wheat from 2011 to 2013, and the result

shows that there is an about 20% efficiency gain on average compared with the direct survey estimates.

2. THE PLANTED CROP ACREAGE ESTIMATION PROJECT

As a major industry in the USA, agriculture is a source of livelihood to millions of farmers, and it is a vital contributor to the global food security. Timely and reliable information about planted crop acreages and productions is very important in that it enables planners and policy makers to propose appropriate strategies for the storage, distribution, and trade of agricultural products. The same information also has a significant impact on farmers through its influence on food prices and crop insurance policies. In the USA, the National Agricultural Statistics Service (NASS) is the federal statistical agency of the US Department of Agriculture (USDA) and is responsible for providing such information, including the estimated planted acreages for different crops on the state level, to the public.

Official crop acreage estimates published by NASS are vetted by the Agricultural Statistics Board, a group of commodity experts who review and assimilate the available survey data and auxiliary information to produce the official estimates. NASS is interested in pursuing model-based estimation strategies for combining information in order to provide accurate crop acreage estimates as well as their uncertainties, and this is the main objective of this project. The methodology we developed in this paper is one of the solutions we provide to address such a practical problem.

In our collaboration with NASS, three sources of information about the planted crop acreage are available. The main source is the June Area Survey (JAS), a national survey conducted annually by NASS to obtain state-level estimates of the planted acreages for various commodity crops. It is an area sample of about 11,000 segments (parcels of land which average approximately 1 square mile in area) selected by a two-stage sampling design. In the first two weeks of June, farm operators who use the land in those segments for agricultural production are interviewed about crops they plant, and the state-level official estimates of planted crop acreages are released later that month in the Acreage Report (United States Department of Agriculture 2015). In addition to JAS data, we have two sources of auxiliary information. One source comes from administrative records collected by the USDA's Farm Service Agency (FSA), which is responsible for the local administration of federal farm programs. Farmers who want to participate in certain federal farm programs will register with FSA and provide the tract-level information, including the planted crop acreage each year. Since the participation in federal agricultural programs is voluntary, not all farmers register with FSA, and those who produce program crops and anticipate government subsidies are more likely to sign up. The other auxiliary information is from NASS's cropland data layer (CDL), a geo-reference crop-specific land cover data layer (Boryan et al. 2011). It has a ground resolution of 30 m and is produced using satellite imageries collected during the current growing season. The primary imagery source is the Indian Remote Sensing IRS-P6 Advanced Wide Field Sensor (AWiFS), which is supplemented by imageries from the Landsat 5 TM and/or Landsat 7 ETM+. Data from FSA and the USDA National Land Cover Dataset 2001 (NLCD 2001) are used as the training and validation dataset, and a

decision tree algorithm C5 is used for classification. The accuracies of large area row crops are between 80 and 90%, but the ones for other crops can be much lower.

3. MODEL SETUP

Direct survey estimates can be improved by a statistical model incorporating auxiliary information. If a model is built by treating individuals as the analysis units, it is called a unit-level model. If a model is constructed using small areas as the analysis units, it is called an area-level model. The unit-level model approach was first considered by Battese et al. (1988) and extended by You and Rao (2002) who developed a pseudo-EBLUP estimator. However, the unit-level model is not applicable when auxiliary variables are not available at the individual level. In many practical situations including our planted crop acreage project, auxiliary information may not be available for each individual. Even if it is available, matching auxiliary information is often difficult or practically impossible at the individual level. In this paper, we consider a hierarchical area-level approach. To be specific, a level-one model is used to reflect the heterogeneity among areas, and a level-two model is used to borrow strength across areas.

The level-one model is specified within areas, and we need to choose the area unit for the level-one model. One choice is the county, which is a natural administrative unit. However, a preliminary analysis indicates that the number of sample units in each county is not large enough to provide reliable results. Instead, we aggregate individual information to a district level, which is an officially predefined grouping of neighboring counties within the same state; more details about the districts are available in the USDA repository https://www.nass.usda.gov/Charts_and_Maps/Crops_County/boundary_maps/indexpdf.php. At the district level, we have more reliable JAS estimates, and there are enough districts to estimate parameters for the level-one model. Note that each state is partitioned into several districts, and information is available for each district. The area unit for the level-two model is chosen to be state, which enables us to provide state-level predictors.

Let $h \in \{1, \ldots, H\}$ be the state index, H be the number of states, and i be the district index within each state. The number of districts in state h is denoted as n_h, and it ranges from 6 to 9 for most states. Let Y_{hi} be the true planted acreage of a specific commodity crop in the ith district of state h, and it is the district-level quantity we are interested in. Denote $\hat{Y}_{hi}$ to be the design-unbiased estimate of Y_{hi} from JAS, and $\hat{V}_{hi}$ to be its design-unbiased variance estimator. Denote $\mathbf{X}_{hi} = (X_{hi1}, X_{hi2})^T$ to be the auxiliary information, where X_{hi1} is the planted crop acreage estimate from FSA, X_{2ih} is the one from CDL, and $\boldsymbol{A}^T$ is the transpose of a matrix $\boldsymbol{A}$. We assume that $\hat{Y}_{hi}$ is conditionally independent of $\mathbf{X}_{hi}$ given Y_{hi}, so we treat $\hat{Y}_{hi}$ as a surrogate variable for Y_{hi}. Furthermore, we assume that the observations for different districts are independent with each other since we do not have enough districts to study the spatial dependence structure within each state. Our goal is to obtain the best predictor of $Y_h = \sum_{i=1}^{n_h} Y_{hi}$ for state h by combining auxiliary information.

By incorporating $\mathbf{X}_{hi}$, we first construct a structural error model (Fay and Herriot 1979),

$$Y_{hi} \sim f_1(Y_{hi} \mid \mathbf{X}_{hi}; \boldsymbol{\theta}_h), \tag{1}$$

for some model $f_1(\cdot)$ known up to a state-specific parameter $\boldsymbol{\theta}_h$. To borrow strength from other states, we assume

$$\boldsymbol{\theta}_h \sim f_2(\boldsymbol{\theta}_h \mid Z_h; \boldsymbol{\zeta}) \tag{2}$$

for some parametric model $f_2(\cdot)$ known up to a parameter $\boldsymbol{\zeta}$, where Z_h is a predefined state-specific covariate indicating the similarity among states. Thus, model (2) is based on groups of similar states. We treat (1) as the level-one model (within-state model) and (2) as the level-two model (between-state model). The state-specific level-one model is helpful to handle the heterogeneity among states, and the level-two model is used to borrow strength from observations outside the state. The two-level model is very useful in describing the hierarchical structure of the data. In addition to structural error models, we use the following sampling error model to incorporate the design-unbiased estimate $\hat{Y}_{hi}$, that is,

$$\hat{Y}_{hi} \sim g_1(\hat{Y}_{hi} \mid Y_{hi})$$

for a distribution $g_1(\cdot)$. In many cases, we can assume that $\hat{Y}_{hi}$ follows from a normal distribution with mean Y_{hi} and variance $\hat{V}_{hi}$.

In our particular application, the level-one model in (1) is not necessarily parametric. Specifically, for each district, an accurate measure of total cultivated acres, denoted as M_{hi}, is available, and it is the upper bound of our planted crop acreage estimate. In order to incorporate such information, let $\bar{Y}_{hi} = Y_{hi}/M_{hi}$ be the *proportion* of the planted crop acreage, and the ones for FSA and CDL estimates are defined in a similar manner. Since $\bar{Y}_{hi}$ takes value from 0 to 1, we consider the following level-one model, that is,

$$\bar{Y}_{hi} = p(\beta_{h0} + \beta_{h1}\bar{X}_{hi1} + \beta_{h2}\bar{X}_{hi2}) + e_{hi}, \tag{3}$$

where $p(x) = \{1 + \exp(-x)\}^{-1}$, $E(e_{hi}|p_{hi}) = 0$, $\text{var}(e_{hi}|p_{hi}) = \psi_h p_{hi}(1 - p_{hi})$, $\text{var}(e_{hi}|p_{hi})$ is the conditional variance of e_{hi} given p_{hi}, ψ_h is a state-specific parameter explaining the over-dispersion among states, and $p_{hi} = p(\beta_{h0} + \beta_{h1}\bar{X}_{hi1} + \beta_{h2}\bar{X}_{hi2})$. Berg and Fuller (2014) also considered model (3) in a single-level model approach. For the level-two model, we use

$$\boldsymbol{\theta}_h \sim N(\boldsymbol{\theta}^{(k)}, \boldsymbol{\Sigma}^{(k)}) \text{ for } h \in U^{(k)}, \tag{4}$$

where $\boldsymbol{\theta}_h = (\beta_{h0}, \beta_{h1}, \beta_{h2})^T$, $N(\boldsymbol{\theta}, \boldsymbol{\Sigma})$ is a normal distribution with mean $\boldsymbol{\theta}$ and covariance $\boldsymbol{\Sigma}$, $\{U^{(k)} : k = 1, \ldots, K\}$ is a prespecified partition of states based on a certain criterion, which serves to group similar states together, and K is the number of groups. This normal assumption for the level-two model is often used in practice. Note that we do not specify a level-two model for the variance parameter ϕ_h. The sampling error model for $\hat{Y}_{hi}$ is

$$\hat{\bar{Y}}_{hi} = \bar{Y}_{hi} + u_{hi}, \quad u_{hi} \sim N(0, \hat{v}_{hi}), \tag{5}$$

where $\hat{v}_{hi} = \hat{V}_{hi}/M_{hi}^2$.

Table 1. Probabilistic structure for the two-level model.

Model	Data	Parameter	Latent variable
Level-one	$\hat{\mathbf{Y}}_h = (\hat{Y}_{h1}, \ldots, \hat{Y}_{hn_h})$	$\boldsymbol{\theta}_h$	$\mathbf{Y}_h = (Y_{h1}, \ldots, Y_{hn_h})$
Level-two	$\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\theta}}_1, \ldots, \hat{\boldsymbol{\theta}}_H)$	$\boldsymbol{\zeta}$	$\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_H)$

4. PARAMETER ESTIMATION

We now discuss the parameter estimation procedure for the proposed two-level model. The estimation procedure is carried out for each state group independently, and we omit Z_h explicitly without loss of generality. Note that there are two types of parameters, that is, $\boldsymbol{\theta}_h$ in the level-one model and $\boldsymbol{\zeta}$ in the level-two model. The true planted crop acreage $\mathbf{Y}_h = (Y_{h1}, \ldots, Y_{hn_h})^T$ is regarded as the latent variable for the level-one model, and $\boldsymbol{\theta}_h$ is the latent variable for the level-two model.

In each level, there are three components, that is, observed data, parameter, and latent variables, and Table 1 presents a summary of the probabilistic structure of these three components. For each level, there are two models involved: one is the structural error model $f(\text{latent} \mid \text{parameter})$, and the other is the sampling error model $g(\text{data} \mid \text{latent})$, which is assumed to be known.

For the case where a distribution is assumed for the structural error model $f(\text{latent} \mid \text{parameter})$, the prediction model for the latent variable is

$$p(y \mid \hat{y}; \theta) = \frac{f(y \mid \theta) g(\hat{y} \mid y)}{\int f(y \mid \theta) g(\hat{y} \mid y) \mathrm{d}y}, \tag{6}$$

where $p(y \mid \hat{y}; \theta)$ is the density function of y given $\hat{y}$ with θ being the parameter and $\hat{y}$, y, and θ denote data, latent variable, and parameter, respectively. Thus, we can use the following learning algorithm to estimate parameters for each level. That is,

Step 1 *Summarization* Obtain the sampling error model for observations of latent variables.

Step 2 *Combination* Find a prediction model for latent variables by combining the sampling error model and the structural error model, and Bayes formula is used as in (6).

Step 3 *Learning* Estimate the parameters.

The proposed estimation procedure can be called a two-level learning algorithm since the learning algorithm will be separately applied at each level.

In the level-one model, we treat $\boldsymbol{\theta}_h$ to be fixed and obtain the best predictor of $\boldsymbol{\theta}_h$ using the proposed learning algorithm. To be more specific, the sampling distribution of the survey estimates $\hat{\mathbf{Y}}_h = (\hat{Y}_{h1}, \ldots, \hat{Y}_{hn_h})^T$ is set to be $g_1(\hat{\mathbf{Y}}_h \mid \mathbf{Y}_h)$, and the prediction model for Y_{hi} is given by

$$p(Y_{hi} \mid X_{hi}, \hat{Y}_{hi}; \boldsymbol{\theta}_h) = \frac{f_1(Y_{hi} \mid \mathbf{X}_{hi}; \boldsymbol{\theta}_h) g_1(\hat{Y}_{hi} \mid Y_{hi})}{\int f_1(Y_{hi} \mid \mathbf{X}_{hi}; \boldsymbol{\theta}_h) g_1(\hat{Y}_{hi} \mid Y_{hi}) \mathrm{d}Y_{hi}}. \tag{7}$$

If $f_1(Y_{hi} \mid \mathbf{X}_{hi}; \boldsymbol{\theta}_h)$ is a normal distribution with mean $\mathbf{X}_{hi}^T \boldsymbol{\beta}_h$ and variance σ_{hi}^2, and $g_1(\hat{Y}_{hi} \mid Y_{hi})$ is also a normal distribution with mean Y_{hi} and variance $\hat{V}_{hi}$, model (7) reduces to

$$Y_{hi} \mid (\mathbf{X}_{hi}, \hat{Y}_{hi}, \boldsymbol{\theta}_h) \sim N\left[c_{hi}\hat{Y}_{hi} + (1 - c_{hi})\mathbf{X}_{hi}^T \boldsymbol{\beta}_h, c_{hi}\hat{V}_{hi}\right],$$

where $c_{hi} = \sigma_{hi}^2/(\hat{V}_{hi} + \sigma_{hi}^2)$. In the learning step, the following EM algorithm can be used.

1. *E-step*: Given the current estimate $\boldsymbol{\theta}_h^{(t)}$, find the conditional distribution of $\mathbf{Y}_h$ given $(\mathbf{X}_h, \hat{\mathbf{Y}}_h)$, where $\mathbf{X}_h$ contains the auxiliary information for state h. That is,

$$p(\mathbf{Y}_h \mid \mathbf{X}_h, \hat{\mathbf{Y}}_h; \boldsymbol{\theta}_h^{(t)}) = \frac{f_1(\mathbf{Y}_h \mid \mathbf{X}_h; \boldsymbol{\theta}_h^{(t)}) g_1(\hat{\mathbf{Y}}_h \mid \mathbf{Y}_h)}{\int f_1(\mathbf{Y}_h \mid \mathbf{X}_h; \boldsymbol{\theta}_h^{(t)}) g_1(\hat{\mathbf{Y}}_h \mid \mathbf{Y}_h) \mathrm{d}\mathbf{Y}_h}.$$

2. *M-step*: Update the estimate of $\boldsymbol{\theta}_h$ by solving

$$E\{\mathbf{S}_1(\boldsymbol{\theta}_h) \mid \mathbf{X}_h, \hat{\mathbf{Y}}_h; \boldsymbol{\theta}_h^{(t)}\} = 0,$$

   where $\mathbf{S}_1(\boldsymbol{\theta}_h) = \partial \log f_1(\mathbf{Y}_h \mid \mathbf{X}_h; \boldsymbol{\theta}_h)/\partial \boldsymbol{\theta}_h$ is the complete-sample score function of $\boldsymbol{\theta}_h$.

3. Repeat E-step and M-step until convergence.

In order to guarantee the convergence property of the EM algorithm, certain restrictions on $f_1(\mathbf{Y}_h \mid \mathbf{X}_h; \boldsymbol{\theta}_h)$ and $g_1(\hat{\mathbf{Y}}_h \mid \mathbf{Y}_h)$ should be satisfied (Wu 1983). Once $\hat{\boldsymbol{\theta}}_h$ is obtained from the above EM algorithm, we also need to obtain the covariance matrix of $\hat{\boldsymbol{\theta}}_h$, denoted as $\hat{\boldsymbol{V}}_h$, and it can be derived using the Louis formula (Louis 1982), that is,

$$\begin{aligned}\mathbf{I}_{obs}(\boldsymbol{\theta}_h) = {} & E\{\mathbf{I}_{com}(\boldsymbol{\theta}_h) \mid \mathbf{X}_h, \hat{\mathbf{Y}}_h; \boldsymbol{\theta}_h\} \\ & + E\{\mathbf{S}_1(\boldsymbol{\theta}_h)^{\otimes 2} \mid \mathbf{X}_h, \hat{\mathbf{Y}}_h; \boldsymbol{\theta}_h\} - \left[E\{\mathbf{S}_1(\boldsymbol{\theta}_h) \mid \mathbf{X}_h, \hat{\mathbf{Y}}_h; \boldsymbol{\theta}_h\}\right]^{\otimes 2},\end{aligned}$$

where $\mathbf{I}_{com}(\boldsymbol{\theta}_h) = -\partial \mathbf{S}_1(\boldsymbol{\theta}_h)/\partial \boldsymbol{\theta}_h^T$ and $\boldsymbol{A}^{\otimes 2} = \boldsymbol{A}\boldsymbol{A}^T$. There are several ways to obtain $\mathbf{I}_{obs}(\boldsymbol{\theta}_h)$; see Kim and Shao (2013) for details.

However, for the case where no parametric model is assumed for the structural error model f(latent | parameter), the EM algorithm cannot be used in the learning step. Specific for the NASS project, by combining (3) with (5), we have

$$\hat{\bar{Y}}_{hi} = p(\beta_{h0} + \beta_{h1}\bar{X}_{hi1} + \beta_{h2}\bar{X}_{hi2}) + e_{hi} + u_{hi},$$

where $E(e_{hi} + u_{hi}) = 0$, $\mathrm{var}(e_{hi} + u_{hi} \mid p_{hi}) = \psi_h p_{hi}(1 - p_{hi}) + \hat{v}_{hi}$, and recall that $\hat{v}_{hi} = \hat{V}_{hi}/M_{hi}^2$. We can use a quasi-likelihood method to estimate $\boldsymbol{\theta}_h$ and ψ_h; see "Appendix A" in the Supplementary Material for details.

For the parameter estimation with respect to the level-two model, use the following learning algorithm.

Step 1 *Summarization*: Assume the sampling error model to be

$$\hat{\boldsymbol{\theta}}_h \mid \boldsymbol{\theta}_h \sim N(\boldsymbol{\theta}_h, \hat{\boldsymbol{V}}_h), \tag{8}$$

where $\hat{\boldsymbol{\theta}}_h$ and $\hat{\boldsymbol{V}}_h$ are obtained from the learning algorithm for the level-one model. This is a commonly used assumption in practice.

Step 2 *Combination* The prediction model for $\boldsymbol{\theta}_h$ is

$$p(\boldsymbol{\theta}_h \mid \hat{\boldsymbol{\theta}}_h; \boldsymbol{\zeta}) = \frac{f_2(\boldsymbol{\theta}_h; \boldsymbol{\zeta}) g_2(\hat{\boldsymbol{\theta}}_h \mid \boldsymbol{\theta}_h)}{\int f_2(\boldsymbol{\theta}_h; \boldsymbol{\zeta}) g_2(\hat{\boldsymbol{\theta}}_h \mid \boldsymbol{\theta}_h) \mathrm{d}\boldsymbol{\theta}_h}, \tag{9}$$

where $g_2(\hat{\boldsymbol{\theta}}_h \mid \boldsymbol{\theta}_h)$ is the normal distribution shown in (8).

Step 3 *Learning*: The parameter is estimated using the EM algorithm as we assume a normal distribution for $f_2(\boldsymbol{\theta}_h; \boldsymbol{\zeta})$. That is,

$$\hat{\boldsymbol{\zeta}}^{(t+1)} = \arg\max_{\boldsymbol{\zeta} \in \boldsymbol{Z}} Q(\boldsymbol{\zeta} \mid \hat{\boldsymbol{\zeta}}^{(t)}),$$

where $\boldsymbol{Z}$ is the feasible region for $\boldsymbol{\zeta}$, and $Q(\boldsymbol{\zeta} \mid \hat{\boldsymbol{\zeta}}^{(t)}) = \sum_{h=1}^{H} E\{\log f_2(\boldsymbol{\theta}_h; \boldsymbol{\zeta}) \mid \hat{\boldsymbol{\theta}}_h; \hat{\boldsymbol{\zeta}}^{(t)}\}$.

Based on the model assumptions, we have

$$\boldsymbol{\theta}_h \mid (\hat{\boldsymbol{\theta}}_h, \boldsymbol{\zeta}) \sim N\left(\boldsymbol{\theta}_h^*, \boldsymbol{V}_h^*\right)$$

with $\hat{\boldsymbol{\theta}}_h^* = \left(\hat{\boldsymbol{V}}_h^{-1} + \boldsymbol{\Sigma}^{-1}\right)^{-1} \left(\hat{\boldsymbol{V}}_h^{-1} \hat{\boldsymbol{\theta}}_h + \boldsymbol{\Sigma}^{-1} \boldsymbol{\theta}\right)$ and $\boldsymbol{V}_h^* = \left(\hat{\boldsymbol{V}}_h^{-1} + \boldsymbol{\Sigma}^{-1}\right)^{-1}$. The convergence of $\{\boldsymbol{\zeta}^{(t)}\}$ is guaranteed since it can be shown that $Q(\boldsymbol{\zeta} \mid \hat{\boldsymbol{\zeta}})$ has finite stationary points, and $f_2(\boldsymbol{\theta}_h; \boldsymbol{\zeta})$ belongs to the exponential family (Wu 1983).

Figure 1 shows a graphical summary for the two-level learning algorithm using a directed acyclic graph. The nodes in the bottom part $(\hat{Y}_{hi}, \mathbf{X}_{hi}, Z_h)$ are the actual observations, and the ones in the upper part $(Y_{hi}, \boldsymbol{\theta}_h)$ are latent variables that we want to predict. The latent variables are used to combine the observed information through statistical models. By the learning procedure with respect to the level-one model, we can estimate $\hat{\boldsymbol{\theta}}_h$ and its covariance estimate $\hat{\boldsymbol{V}}_h$. Once they are obtained, we can use them in the summarization step of learning procedure for the level-two model and apply the EM algorithm to estimate $\boldsymbol{\zeta}$. When all parameters are estimated, we can get the best predictors, which will be discussed in the next section.

5. PREDICTION

Once the parameters are estimated, we can obtain the best predictor of $Y_h = \sum_i Y_{hi}$. Under the model setup in Sect. 3, given $\boldsymbol{\theta}_h$, the best predictor of Y_{hi} is

$$\tilde{Y}_{hi}^*(\boldsymbol{\theta}_h) = E(Y_{hi} \mid \mathbf{X}_{hi}, \hat{Y}_{hi}; \boldsymbol{\theta}_h), \tag{10}$$

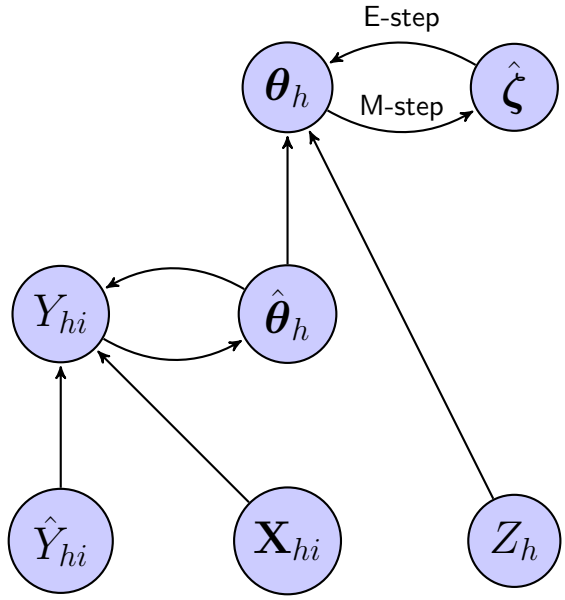

Figure 1. Two-level learning algorithm.

where the conditional expectation is taken with respect to the prediction model in (7). Since the conditional expectation in (10) involving a latent variable $\boldsymbol{\theta}_h$, we use

$$\tilde{Y}_{hi}^{**} = E\{E(Y_{hi} \mid \mathbf{X}_{hi}, \hat{Y}_{hi}; \boldsymbol{\theta}_h) \mid \hat{\boldsymbol{\theta}}_h; \boldsymbol{\zeta}\},$$

where the first conditional expectation is with respect to the second-level prediction model in (9). Thus, the best predictor of Y_{hi} under the two-level model and mean squared error loss is

$$\hat{Y}_{hi}^{**} = E\{E(Y_{hi} \mid \mathbf{X}_{hi}, \hat{Y}_{hi}; \boldsymbol{\theta}_h) \mid \hat{\boldsymbol{\theta}}_h; \hat{\boldsymbol{\zeta}}\}.$$

Because $\hat{\boldsymbol{\theta}}_h^*$ is used in (10) in place of $\hat{\boldsymbol{\theta}}_h$, the state-level prediction $\hat{Y}_h^{**} = \sum_i \hat{Y}_{hi}^{**}$ borrows strength from observations outside states h.

Specific for the logistic model (3) with a normal distribution assumption for the level-two model (4), the best predictor of $\bar{Y}_{hi}$ is

$$\hat{\bar{Y}}_{hi}^{*} = \hat{c}_{hi}\hat{\bar{Y}}_{hi} + (1 - \hat{c}_{hi})\hat{p}_{hi}^{*}, \tag{11}$$

where $\hat{p}_{hi}^{*} = p(\hat{\beta}_{h0}^{*} + \hat{\beta}_{h1}^{*}\bar{X}_{hi1} + \hat{\beta}_{h2}^{*}\bar{X}_{hi2})$ and $\hat{c}_{hi} = [\hat{\psi}_h \hat{p}_{hi}^{*}(1 - \hat{p}_{hi}^{*})]/[\hat{v}_{hi} + \hat{\psi}_h \hat{p}_{hi}^{*}(1 - \hat{p}_{hi}^{*})]$. Thus, $\hat{Y}_h^{*} = \sum_{i=1}^{m_h} M_{hi}\hat{\bar{Y}}_{hi}^{*}$ is the best predictor of Y_h.

The mean squared prediction error of $\hat{Y}_h^{*}$ is used to assess the uncertainty of this best predictor, and it is defined by

$$\text{MSPE}(\hat{Y}_h^{*}) = E\left\{\left(\hat{Y}_h^{*} - Y_h\right)^2\right\},$$

where the expectation is taken conditional on the observations. A detailed estimation procedure for the mean squared prediction error is given in "Appendix B" in the Supplementary Material.

Remark. By resembling a real case study, Ghosh and Rao (1994) conducted a simulation to compare different small area estimates. Instead of our multi-level model, Torabi and Rao (2014) considered a sub-area-level model, and a simulation study is used to test its performance. The proposed two-level model differs from the sub-area-level model considered by Torabi and Rao (2014) in the following aspects. Instead of specifying the regression parameter to be the same for different areas and using an area random effect model, we build a level-two model for all parameters in the level-one model such that we can characterize the heterogeneity among the coefficient parameters. By resembling the NASS planted crop acreage estimation project, a simulation study is carried out to test the performance of the proposed two-level model, and it shows that the proposed estimation procedure works well, and the two-level model can be used to improve the estimates by incorporating auxiliary information; see "Appendix C" in the Supplementary Material for details. Since the estimate of the mean squared prediction error is based on asymptotic results, there may exist some numerical issues when the sample size is small.

6. APPLICATION TO NASS PROJECT

In this section, we apply the two-level model to improve the planted crop acreage estimate by combining information from the JAS, FSA, and CDL. The datasets for districts, FSA and CDL, are available in the USDA repositories https://www.nass.usda.gov/Charts_and_Maps/Crops_County/boundary_maps/indexpdf.php, https://www.fsa.usda.gov/news-room/efoia/electronic-reading-room/frequently-requested-information/crop-acreage-data/ and https://www.nass.usda.gov/Research_and_Science/Cropland/Release/, respectively. State- and district-level direct estimates obtained from the June Area Survey are not publicly available in conformance with NASS disclosure guidelines, but a surrogate can be obtained by contacting the authors.

For the NASS planted crop acreage estimation project, we classify states into two crop-specific groups, that is, a "major" group and a "minor" group. A state is classified to the "major" group of a commodity crop if the median of district-level planted crop acreages is greater than 150,000 acres, and CDL classified acreages are used to make this determination. Otherwise, the state is in the "minor" group. For a specific crop, let $Z_h = 1$ if state h belongs to the "major" group, and $Z_h = 0$ otherwise. Then, Z_h is a predefined covariate in model (2), and it determines the sets $U^{(1)}$ and $U^{(2)}$ in model (4). Since the group is crop specific, it is possible for a state to be classified to the "major" group with respect to one commodity (e.g., corn) and being classified to the "minor" group of another (e.g., winter wheat).

We use (3) as the level-one model, (4) as the level-two model, and (5) and (8) as the sampling error models. Since the collinearity between the FSA and CDL estimates may cause numerical issues, we also apply the proposed model by using FSA or CDL as the single covariate. Before making inference, we check the assumption made

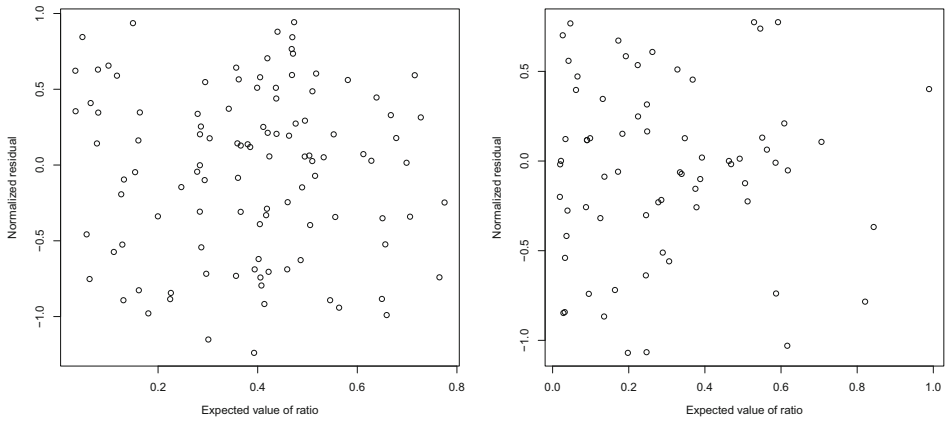

Figure 2. Relationship between the normalized residuals and the fitted values from the level-one model for soybeans in 2013. The result on the left panel is for 'major' crop states, and the one on the right panel for 'minor' crop states.

for the error term in the level-one model based on the normalized residuals $\hat{e}_{hi} = (\hat{\bar{Y}}_{hi} - \hat{p}_{hi})/\sqrt{\hat{v}_{hi} + \hat{\psi}_h \hat{p}_{hi}(1 - \hat{p}_{hi})}$, where $\hat{p}_{hi} = p(\hat{\beta}_{h0} + \hat{\beta}_{h1}\bar{X}_{hi1} + \hat{\beta}_{h2}\bar{X}_{hi2})$, $\hat{\beta}_{h0}$, $\hat{\beta}_{h1}$, $\hat{\beta}_{h2}$, and $\hat{\psi}_h$ are the estimators by the learning algorithm for the level-one model. Figure 2 shows the relationship between $\hat{e}_{hi}$ and $\hat{p}_{hi}$ for the soybeans in 2013. No obvious trend or pattern is apparent, and the variance can be regarded as constant. Thus, the assumptions made in the level-one model are valid. Similar results hold for other cases.

Different estimates are compared by averaging the relative efficiencies on the state level. That is,

$$\text{ARE} = \frac{1}{H}\sum_{h=1}^{H} \frac{\text{MSPE}(\hat{Y}_h^*)}{\sum_{i=1}^{n_h} \hat{V}_{hi}},$$

where $\sum_{i=1}^{n_h} \hat{V}_{hi}$ is the variance of $\hat{Y}_h = \sum_{i=1}^{n_h} \hat{Y}_{hi}$ according to the independence assumption for $\hat{\mathbf{Y}}_h$. Table 2 shows the results of the average relative efficiencies for corn, soybean, and winter wheat, which are three major row crops in the USA. The estimates by the proposed two-level model are more efficient than direct survey estimates for all cases in the sense that the average relative efficiencies are smaller than 1. The estimates using both FSA and CDL as covariates have smaller estimated mean square prediction errors on average than the ones using a single covariate, but there may exist some numerical issues due to the collinearity problem. On general, there is about 20% improvement in efficiency by using the proposed two-level model compared with the direct survey estimate.

Figure 3 demonstrates a comparison among different estimates of the planted crop acreage for states in the "major" group of soybeans in 2013. For each state, we present six estimates: JAS, FSA, CDL, and three estimators by the proposed two-level model using both FSA and CDL as covariates, FSA as the single covariate and CDL as the single covariate, respectively. The 1.96 standard errors (square roots of the estimated mean squared prediction error) are also shown for the JAS estimate and the ones by the proposed two-level model, and the corresponding interval can be used to approximate the 95% confidence interval based on

Table 2. Summaries of the average relative efficiencies for different estimation methods.

Year	Crop	Group	Both	FSA	CDL
2011	Corn	Major	0.73	0.76	0.79
		Minor	0.64	0.72	0.72
	Soybeans	Major	0.76	0.75	0.80
		Minor	0.71	0.72	0.73
	Winter Wheat	Major	0.62	0.79	0.74
		Minor	0.75	0.82	0.83
2012	Corn	Major	0.77	0.82	0.79
		Minor	0.68	0.75	0.77
	Soybeans	Major	0.79	0.86	0.87
		Minor	N.A.[a]	0.72	0.64
	Winter Wheat	Major	0.74	0.80	0.80
		Minor	0.63	0.70	0.69
2013	Corn	Major	0.74	0.78	0.81
		Minor	0.56	0.68	0.68
	Soybeans	Major	0.64	0.71	0.74
		Minor	0.57	0.62	0.67
	Winter Wheat	Major	N.A.[a]	0.74	0.84
		Minor	0.50	0.66	0.69

Both shows the results by using both FSA and CDL in the proposed two-level model, "FSA" the ones using only FSA as the single covariate, and "CDL" the ones using CDL as the single covariate.

[a]The two-level model encounters numerical problems due to the collinearity between the FSA and CDL estimates within some states. Thus, no result is available in this case.

the asymptotic normality. The estimates by the proposed two-level model are more efficient than the direct survey estimate since the estimated mean squared prediction error is smaller than the variance of JAS estimate. Besides, the estimated mean squared prediction error of the estimates by using both FSA and CDL as covariates is similar or smaller than the ones by using a single covariate for most states.

7. CONCLUDING REMARKS

We consider a hierarchical area-level approach for a small area estimation problem by combining information from three sources. A two-level model is developed to characterize the state-specific heterogeneity as well as to borrow information from other states. To estimate parameters in the two-level model, we propose a frequentist learning algorithm, and it can be naturally extended for general multi-level models. If the distribution assumption is made for the level-one model, a Bayesian method can also be applied. The proposed two-level model is applied to estimate the state-level planted crop acreages for the NASS project, and it shows improvement over the direct survey estimates since the estimated mean squared prediction error is smaller than the design variance on general.

The proposed method is based on the Fay–Herriot model approach. Instead of a Fay–Herriot model approach, one can also consider a measurement error model approach, where the structural model is made to regress $(\mathbf{X}_{hi}, \hat{Y}_{hi})$ on Y_{hi} (Kim et al. 2015). Extension of the proposed method to two-level measurement error model approach will be a topic of

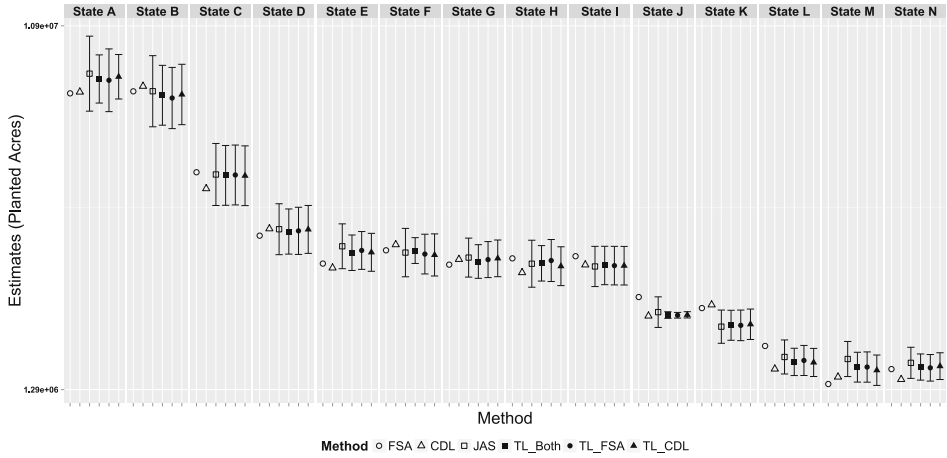

Figure 3. State-level estimates of the planted crop acreage of the 2013 soybeans for the "major" crop group. For each state, we compare six estimates, including FSA, CDL, JAS and the ones by the proposed two-level model using both FSA and CDL as covariate (TL_Both), FSA as the single covariate (TL_FSA) and CDL as the single covariate (TL_CDL). The vertical line shows the 1.96 standard errors for JAS, and 1.96 square roots of mean squared prediction error for the estimates by the two-level model.

future research. We did not consider the spatial dependence structure of the observations among different districts in the same state due to the limited information. For the case where there are enough district-level observations for each state, we could either use a parametric spatial model (Cressie 2015) or a nonparametric one (Lahiri and Zhu 2006) to improve the level-one model, and this could be another research topic in the future.

ACKNOWLEDGEMENTS

We are grateful to three referees and the Associate Editor for the constructive comments. This research was supported by the National Agricultural Statistics Service of the US Department of Agriculture.

[Received February 2017. Accepted March 2018. Published Online April 2018.]

REFERENCES

Battese, G. E., Harter, R. M. and Fuller, W. A. (1988). An error-components model for prediction of county crop areas using survey and satellite data, *Journal of the American Statistical Association* **83**: 28–36.

Berg, E. J. and Fuller, W. A. (2014). Small area prediction of proportions with applications to the canadian labour force survey, *Journal of Survey Statistics and Methodology* **2**: 227–256.

Boryan, C., Yang, Z., Mueller, R. and Craig, M. (2011). Monitoring us agriculture: the US department of agriculture, national agricultural statistics service, cropland data layer program, *Geocarto International* **26**: 341–358.

Cressie, N. (2015). *Statistics for Spatial Data*, revised edn, John Wiley & Sons, New York.

Datta, G., Ghosh, M. et al. (2012). Small area shrinkage estimation, *Statistical Science* **27**: 95–114.

Datta, G. S. (2009). Model-based approach to small area estimation, *Handbook of Statistics* **29**: 251–288.

Datta, G. S. and Ghosh, M. (1991). Bayesian prediction in linear models: Applications to small area estimation, *The Annals of Statistics* **19**: 1748–1770.

Deming, W. E. and Stephan, F. F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known, *The Annals of Mathematical Statistics* **11**: 427–444.

Dever, J. A. and Valliant, R. (2010). A comparison of variance estimators for poststratification to estimated control totals, *Survey Methodology* **36**: 45–56.

Elliott, M. R. and Davis, W. W. (2005). Obtaining cancer risk factor prevalence estimates in small areas: combining data from two surveys, *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **54**: 595–609.

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.

Ghosh, M., Natarajan, K., Stroud, T. and Carlin, B. P. (1998). Generalized linear models for small-area estimation, *Journal of the American Statistical Association* **93**: 273–282.

Ghosh, M. and Rao, J. N. K. (1994). Small area estimation: an appraisal, *Statistical science* **9**: 55–76.

Hidiroglou, M. (2001). Double sampling, *Survey methodology* **27**: 143–154.

Kim, J. K. and Park, M. (2010). Calibration estimation in survey sampling, *International Statistical Review* **78**: 21–39.

Kim, J. K., Park, S. and Kim, S. Y. (2015). Small area estimation combining information from several sources, *Survey Methodology* **41**: 21–36.

Kim, J. K. and Rao, J. N. K. (2012). Combining data from two independent surveys: a model-assisted approach, *Biometrika* **99**: 85–100.

Kim, J. K. and Shao, J. (2013). *Statistical Methods for Handling Incomplete Data*, CRC Press, Florida.

Lahiri, S. N. and Zhu, J. (2006). Resampling methods for spatial regression models under a class of stochastic designs, *The Annals of Statistics* **34**: 1774–1813.

Legg, J. C. and Fuller, W. A. (2009). Two-phase sampling, *Handbook of statistics* **29**: 55–70.

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm, *Journal of the Royal Statistical Society. Series B (Methodological)* **44**: 226–233.

Manzi, G., Spiegelhalter, D. J., Turner, R. M., Flowers, J. and Thompson, S. G. (2011). Modelling bias in combining small area prevalence estimates from multiple surveys, *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **174**: 31–50.

Merkouris, T. (2004). Combining independent regression estimators from multiple surveys, *Journal of the American Statistical Association* **99**: 1131–1139.

Merkouris, T. (2010). Combining information from multiple surveys by using regression for efficient small domain estimation, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **72**: 27–48.

Pfeffermann, D. (2002). Small area estimation: New developments and directions, *International Statistical Review/Revue Internationale de Statistique* **70**: 125–143.

Raghunathan, T. E., Xie, D., Schenker, N., Parsons, V. L., Davis, W. W., Dodd, K. W. and Feuer, E. J. (2007). Combining information from two surveys to estimate county-level prevalence rates of cancer risk factors and screening, *Journal of the American Statistical Association* **102**: 474–486.

Rao, J. N. K. and Molina, I. (2015). *Small Area Estimation*, second edn, Wiley Online Library, New Jersey.

Renssen, R. H. and Nieuwenbroek, N. J. (1997). Aligning estimates for common variables in two or more sample surveys, *Journal of the American Statistical Association* **92**: 368–374.

Tam, S.-M. and Clarke, F. (2015). Big data, official statistics and some initiatives by the Australian Bureau of Statistics, *International Statistical Review* **83**: 436–448.

Torabi, M. and Rao, J. N. K. (2008). Small area estimation under a two-level model, *Survey Methodology* **34**: 11–17.

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.

United States Department of Agriculture (2015). June area survey, Website. Last checked: October 15, 2015.

Wu, C. J. (1983). On the convergence properties of the EM algorithm, *The Annals of statistics* **11**: 95–103.

Wu, C. and Lu, W. W. (2016). Calibration weighting methods for complex surveys, *International Statistical Review* **84**: 79–98.

You, Y. and Rao, J. N. K. (2002). A pseudo-empirical best linear unbiased prediction approach to small area estimation using survey weights, *The Canadian Journal of Statistics/La Revue Canadienne de Statistique* **30**: 431–439.

Zieschang, K. D. (1990). Sample weighting methods and estimation of totals in the consumer expenditure survey, *Journal of the American Statistical Association* **85**: 986–1001.