Mixture Model and Its Application

Yang Cheng[1], Gauri Datta[2], Lu Chen[1,3] Balgobin Nandram[1,4], and Ruiyi Zhang[1,3]

Abstract
Federal Statistical Agencies are required to produce estimates of subpopulation parameters. Because the number of sample survey observations within a subpopulation tends to become small as the size of the subpopulation decreases, traditional design-based estimates based on only sample survey data from subpopulations are often unreliable. For over forty years, the Fay-Herriot model has been widely used to produce reliable small area statistics. This model develops predictions of small areas of interest based on a linear regression of the response of interest on auxiliary variables. For the Fay-Herriot model, the response variable is assumed to be normally distributed, and the random effects associated with various levels of geography and sampling errors are assumed to be independent, normally distributed random variables with a mean of zero and an unknown variance. The Fay-Herriot model is sensitive to outliers because the outliers may result in overestimation of the model variance. In this talk, we propose a new robust estimation approach to estimate small area populations. The robustness property is achieved by replacing the standard normality assumption of the model errors by a mixture of two normal distributions with different variances, making this mixture model less sensitive to outliers. Finally, we compare the estimates from the proposed mixture model to alternative existing methods using a data set from the Cash Rents Survey conducted by the United States Department of Agriculture's (USDA's) National Agricultural Statistics Service (NASS).

Key Words: Mixture model, Small area estimation, Fay-Herriot model, Cash rents survey

1. Introduction

The motivation behind the mixture model is that the available data may include unobserved subgroups and, by incorporating such structure in the model, we could obtain more accurate predictions. In statistics, a mixture model is a probabilistic model for representing the presence of subpopulations within an overall population, without requiring that an observed data set should identify the subpopulation to which an individual observation belongs.

Sample surveys are used to collect useful data from the population and estimate various population characteristics from the sample. The sample should be representative of the population, ideally with participants selected at random from the target population.

[1] National Agricultural Statistics Service, 1400 Independence Ave., SW, Washington, DC 20250
[2] University of Georgia, 310 Herty Dr, Athens, GA 30602
[3] National Institute of Statistical Sciences, 1750 K Street, NW, Suite 1100, Washington, DC 2006
[4] Worcester Polytechnic Institute,100 Institute Road, Worcester, MA 01609

Because results tend to vary with samples, it is necessary to quantify the sampling error or variation that exists among estimates from different samples.

In this section, we introduce area-level model for small area problem and briefly review the Fay-Herriot model proposed by Fay and Herriot (1979) in the small area estimation. We propose a new mixture model in Section 2 to relax the model error in the linking equation and write out the Maximum Likelihood Estimation (MLE) of parameters. In Section 3, we develop EM algorithm for the Maximum Likelihood Estimation. We apply proposed mixture model with EM algorithm on the Cash Rents Survey in Section 4. Finally, we discuss future research on the mixture model.

1.1 Small area problem

Surveys are usually designed to have enough samples at the state or national level to produce reliable design-based estimates with a desired level of precision. When estimating areas/domains, a problem may arise because the sample size may not be large enough for reliable design-based estimates. The problem of estimation at such detailed levels with inadequate or no sample is known as the small area estimation (SAE) problem (Rao and Molina, 2005).

Any improvement of the direct small area estimator without collecting additional new data requires certain implicit or explicit modelling assumptions. It is common to encounter situations where a reasonable working model can be found to explain the bulk of the data. However, a handful of data points may not fit the model, adversely affecting estimation of the model parameters and hence the small area parameters. This calls for development of an estimation method that is robust to the occurrence of outliers or misspecification of modelling assumptions.

SAE methods combine and borrow strength from multiple data sources, such as surveys, census, administrative data, and the choice usually depends on the parameter being estimated and data sources available. They yield estimates that remain reliable even when disaggregated at levels, or "small areas", for which the survey was not originally designed to provide reliable estimates.

Consider the basic area model for the small area problem,

$$\begin{cases} y_i = \theta_i + e_i & -\ sampling\ equation \\ \theta_i = x_i^T \beta + v_i & -\ linking\ equation \end{cases} \qquad (1)$$

where y_i is a direct design-based estimate, θ_i is true area parameter, e_i is sampling error, $e_i \sim N(0, D_i)$, $i = 1,2, \cdots n,\ D_i$ is known and positive. y_i and D_i are directly obtained from sample estimates. In the linking equation, x_i is a vector of covariates, and v_i is model error.

1.2 Background on Fay-Herriot model

Fay and Herriot (1979) proposed model (Equation 1) to develop estimates of small area means based on direct survey estimates (y_i) and synthetic regression estimates computed from auxiliary variables (x_i). Their model, which is essentially a mixed linear model, is widely known as the Fay-Herriot model in the small area estimation literature. In the standard Fay-Herriot model, it is assumed that the model errors (v_i's) are identically

distributed with a common normal distribution with mean zero and an unknown variance, that is, $v_i \sim N(0, \sigma^2)$. The model parameters are β and σ^2.

We need to estimate model parameters: (β, σ^2). The best predictor of θ_i under squared error loss is the conditional expectation given by

$$E(\theta_i|y_i) = \gamma_i y_i + (1-\gamma_i)x_i^T\beta$$

where $\gamma_i = \sigma^2/(\sigma^2 + D_i)$ is known as a shrinkage coefficient. Under the area model, the marginal distribution of y_i is $N(x_i^T\beta, \sigma^2 + D_i)$, and β can be estimated for a given σ^2 by generalized least squares (GLS) estimator given by

$$\hat{\beta}_{GLS} = \left(\sum_{i=1}^{n} \frac{x_i x_i^T}{\sigma^2 + D_i}\right)^{-1} \left(\sum_{i=1}^{n} \frac{x_i y_i}{\sigma^2 + D_i}\right)$$

By replacing β with $\hat{\beta}_{GLS}$, we obtain the following best linear unbiased predictor (BLUP) of θ_i:

$$\tilde{\theta}_i = \gamma_i y_i + (1-\gamma_i)x_i^T\hat{\beta}_{GLS}$$

1.3 Estimating the model parameters

In practice, the random effects variance σ^2 is unknown and should be replaced in γ_i and $\hat{\beta}_{GLS}$ by a sample estimate, which yields the empirical BLUP in the frequentist's framework, or the empirical Bayes estimator in the Bayesian framework. To estimate σ^2, we will consider the following approaches:

- Maximum likelihood estimator based on the marginal distribution of y_i
- Restricted maximum likelihood estimator and moment-type estimators
- Hierarchical Bayes (HB) approach by assigning prior distributions on unknown parameters β and σ^2, and compute a posterior distribution of θ_i
- Log transformed data.

From the Fay-Herriot small area model, we are able to estimate (β, σ^2) as $(\hat{\beta}, \hat{\sigma}^2)$. The two major methods of parameter estimation for process models are maximum likelihood and least squares. Both methods provide parameter estimators that have good properties. However, both maximum likelihood and least squares are sensitive to the presence of outliers.

2. Mixture Model

2.1 Model misspecification or the presence of outliers

The motivation for the proposed model is to develop estimators when the model is misspecified or outliers are present. The Fay-Herriot assumption on the model errors is found to be restrictive in many applications (Lahiri and Rao, 1995; Datta and Lahiri, 1995). To relax the distributional assumption, we assume that the v_i's are identically distributed, and the common distribution is a mixture of two normal distributions. We propose a model by changing the distribution of the model error, v_i, from normal to a mixture of two normal,

$$\begin{cases} v_i|z_i = 1 \sim N(0, \sigma^2) \\ v_i|z_i = 0 \sim N(0, \rho\sigma^2) \\ \quad z_i|p \sim Ber(p) \end{cases} \qquad (2)$$

where the mixture indicator (z_i) is a Bernoulli variable and $0 < \rho < 1$. That is, when $z_i = 1$, the v_i follows the normal distribution with larger variability, which accommodates the outliers.

Now, model parameters have changed from (β, σ^2) in the Fay-Herriot model to $\Phi = \{\beta, \sigma^2, \rho, \text{ p}\}$. We will consider the maximum likelihood approach to estimate Φ from the complete dataset: (Y, Z, V) where Y is estimated from the survey, and (Z, V) are the unobserved latent variables.

2.2 Maximum Likelihood Estimation of Parameters

The complete likelihood function is the joint density of $(Y,\ Z,\ V)$, which is given by

$$L(\phi; Y, V, Z) = \prod_{i=1}^{n} f(y_i|v_i)\, f(v_i|z_i) f(z_i) \qquad (3)$$

where $f(y_i|v_i) = \frac{1}{\sqrt{2\pi D_i}} exp\left[-\frac{(y_i - x_i^T\beta - v_i)^2}{2D_i}\right]$, $f(v_i|z_i) = [N(v_i|0, \sigma^2)]^{z_i}[N(v_i|0, \rho\sigma^2)]^{(1-z_i)}$, and $f(z_i) = p^{z_i}(1-p)^{1-z_i}$.

Then the likelihood function is

$$L(\phi; Y, V, Z) = \prod_{i=1}^{n} \{N\left(y_i|x_i^T\beta + v_i, D_i\right)\left(pN(v_i|0, \sigma^2)\right)^{z_i}\left((1-p)N(v_i|0, \rho\sigma^2)\right)^{1-z_i}\}$$

The log-likelihood function for the complete data is given by

$$l(\phi; Y, V, Z) = \ln[L(\phi; Y, V, Z)]$$

$$= \text{ constant - } \sum_{i=1}^{n} \frac{(y_i - x_i^T\beta - v_i)^2}{2D_i} + \sum_{i=1}^{n} z_i \left(log(p) - \frac{1}{2} log(2\pi\sigma^2) - \frac{v_i^2}{2\sigma^2}\right) + \sum_{i=1}^{n}(1 - z_i)\left(log(1-p) - \frac{1}{2} log(2\pi\rho\sigma^2) - \frac{v_i^2}{2\rho\sigma^2}\right) \qquad (4)$$

3. EM algorithm for Maximum Likelihood Estimation

3.1 EM algorithm

Consider the maximum likelihood approach to estimate model parameters: $\Phi = \{\beta, \sigma^2, \rho, \text{ p}\}$. In mixture models, the likelihood functions are usually too complicated to deal with via standard maximization. The EM algorithm, recommended by Dempster et al. (1977), is effective and popular for maximizing likelihood function for mixture models. In the EM terminology, the observed data $y = (y_1, y_2, \cdots, y_n)^T$ is referred to as the incomplete data. If the unobserved latent variables $z_1, z_2, \cdots, z_n, v_1, v_2, \cdots, v_n$ are available, we can write down the log-likelihood function for the complete data $y_1, y_2, \cdots, y_n, z_1, z_2, \cdots, z_n, v_1, v_2, \cdots, v_n$.

E-step: Given the current parameter $\phi^{(t)}$ and observed data Y, compute the expectation of $l(\phi; Y, V, Z)$ on the joint density function of (V, Z)

$$Q\left(\phi; \phi^{(t)}, Y\right) = E_{(V,Z)}\left(l(\phi; Y, V, Z)|\phi^{(t)}, Y\right) = E_Z\{E_V\left[l(\phi; Y, V, Z)|\phi^{(t)}, Y, Z\right]|\phi^{(t)}, Y\}$$

Note that:

$$v_i|y_i, z_i,\ \phi^{(t)} \sim N\left(\left(\frac{1}{D_i} + \frac{z_i}{\sigma^{2(t)}} + \frac{1-z_i}{\sigma^{2(t)}\rho^{(t)}}\right)^{-1} \frac{y_i - x_i^T\beta^{(t)}}{\mathrm{D_i}}, \left(\frac{1}{D_i} + \frac{z_i}{\sigma^{2(t)}} + \frac{1-z_i}{\sigma^{2(t)}\rho^{(t)}}\right)^{-1}\right).$$

So, we need to calculate $E_{V|Z}\left(v_i\middle|Y, z_i,\ \phi^{(t)}\right)$ and $E_{V|Z}\left(v_i^2\middle|Y, z_i,\ \phi^{(t)}\right)$ for $E_{V|Z}\left[l(\phi; Y, V, Z)|\phi^{(t)}, Y, Z\right]$:

1. $E_{V|Z}\left(v_i\middle|Y, z_i,\ \phi^{(t)}\right) = \left(\frac{1}{D_i} + \frac{z_i}{\sigma^{2(t)}} + \frac{1-z_i}{\sigma^{2(t)}\rho^{(t)}}\right)^{-1} \frac{y_i - x_i^T\beta^{(t)}}{\mathrm{D_i}} \triangleq R_1\left(z_i; y_i, \phi^{(t)}\right)$
2. $E_{V|Z}\left(v_i^2\middle|Y, z_i,\ \phi^{(t)}\right) = \left(\frac{1}{D_i} + \frac{z_i}{\sigma^{2(t)}} + \frac{1-z_i}{\sigma^{2(t)}\rho^{(t)}}\right)^{-1} + \left[E_V\left(v_i\middle|Y, z_i,\ \theta^{(t)}\right)\right]^2 \triangleq R_2(z_i; y_i, \phi^{(t)})$

The conditional probability of Z_i given y_i and $\phi^{(t)}$ is

$$P\left[Z_i = z_i \mid Y = y_i, \phi^{(t)}\right] = \frac{p^{(t)^{z_i}}(1-p^{(t)})^{1-z_i} N\left(y_i \mid x_i^{\mathrm{T}}\beta^{(t)}, D_i + \sigma^{2(t)}\rho^{(t)^{(1-z_i)}}\right)}{p^{(t)} N\left(y_i\middle|x_i^T\beta^{(t)}, D_i + \sigma^{2(t)}\right) + (1-p^{(t)}) N\left(y_i\middle|x_i^T\beta^{(t)}, D_i + \rho^{(t)}\sigma^{2(t)}\right)} = \pi_i\left(z_i; y_i, \phi^{(t)}\right)$$

with $\pi_i\left(0; y_i, \phi^{(t)}\right) + \pi_i\left(1; y_i, \phi^{(t)}\right) = 1$.
Let

$$\begin{aligned} Q\left(\phi; \phi^{(t)}, Y\right) = \text{constant} &- \sum_{i=1}^{n} \frac{\left(y_i - x_i^T\beta\right)^2}{2D_i} + \sum_{i=1}^{n}\left[ln(1-p) - \frac{1}{2}ln(\rho\sigma^2)\right] \\ &+ \sum_{i=1}^{n} \pi_i\left(1; y_i, \phi^{(t)}\right)\left[ln(p) - ln(1-p) + \frac{1}{2}ln(\rho)\right] \\ &- \sum_{i=1}^{n} \frac{\left(y_i - x_i^T\beta\right)^2}{D_i} R_3\left(y_i, \phi^{(t)}\right) \\ &- \frac{1}{2}\sum_{i=1}^{n}\left(\frac{1}{D_i} + \frac{1}{\sigma^2}\right) R_2\left(z_i = 1; y_i, \phi^{(t)}\right)\pi_i\left(1; y_i, \phi^{(t)}\right) \\ &- \frac{1}{2}\sum_{i=1}^{n}\left(\frac{1}{D_i} + \frac{1}{\rho\sigma^2}\right) R_2\left(z_i = 0; y_i, \phi^{(t)}\right)\pi_i\left(0; y_i, \phi^{(t)}\right) \end{aligned}$$

where $R_3\left(y_i, \phi^{(t)}\right) = R_1\left(z_i = 1; y_i, \phi^{(t)}\right)\pi_i\left(1; y_i, \phi^{(t)}\right) + R_1\left(z_i = 0; y_i, \phi^{(t)}\right)\pi_i\left(0; y_i, \phi^{(t)}\right)$.

M-step: Update ϕ by

$$\phi^{(t+1)} = arg\max_{\phi} Q\left(\phi; \phi^{(t)}, Y\right)$$

So, we need to calculate $\beta^{(t+1)}, p^{(t+1)}, {\sigma^2}^{(t+1)}$, $and\ \rho^{(t+1)}$.

1. $\frac{\partial Q(\phi;\phi^{(t)},Y)}{\partial \beta} = 0 \Rightarrow \hat{\beta}^{(t+1)} = \left(\sum_{i=1}^{n} \frac{x_i x_i^T}{D_i}\right)^{-1} \sum_{i=1}^{n} \frac{x_i^T [y_i - R_3(y_i, \phi^{(t)})]}{D_i}$
2. $\frac{\partial Q(\phi;\phi^{(t)},Y)}{\partial p} = 0 \Rightarrow\ \hat{p}^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n} \pi_i\left(1; y_i, \phi^{(t)}\right)$
3. $\begin{cases} \frac{\partial Q(\phi;\phi^{(t)},Y)}{\partial \sigma^2} &= 0 \\ \frac{\partial Q(\phi;\phi^{(t)},Y)}{\partial \rho} &= 0 \end{cases} \Rightarrow \begin{cases} \hat{\sigma}^{2(t+1)} &= \frac{\sum_{i=1}^{n} R_2(z_i=1; y_i, \phi^{(t)}) \pi_i(1; y_i, \phi^{(t)})}{\sum_{i=1}^{n} \pi_i(1; y_i, \phi^{(t)})} \\ \hat{\rho}^{(t+1)} &= \frac{\sum_{i=1}^{n} R_2(z_i=0; y_i, \phi^{(t)}) \pi_i(0; y_i, \phi^{(t)})}{\sum_{i=1}^{n} \pi_i(0; y_i, \phi^{(t)})} \frac{1}{\hat{\sigma}^{2(t+1)}} \end{cases}$

3.2 Empirical best predictor

For EM algorithm calculations, define the initial value, $\phi^{(0)}$, for model parameters Φ. Let $\beta^{(0)} = \hat{\beta}$, $(\sigma^2)^{(0)} = \hat{\sigma}^2$, $p^{(0)} = .5$, and $\rho^{(0)} = .5$, where $\hat{\beta}$ and $\hat{\sigma}^2$ are derived from the Fay-Herriot model.

For known model parameters $\Phi = (\beta, p, \sigma^2, \rho)$, the best predictor of θ_i is

$$\tilde{\theta}_{i,\mathrm{B}}(\phi) = x_i^{\mathrm{T}}\beta + \pi_i(1; y_i, \phi)\frac{\sigma^2}{\sigma^2 + D_i}\left(y_i - x_i^{\mathrm{T}}\beta\right) + \pi_i(0; y_i, \phi)\frac{\rho\sigma^2}{\rho\sigma^2 + D_i}\left(y_i - x_i^{\mathrm{T}}\beta\right)$$

The empirical best predictor of θ_i is

$$\hat{\theta}_{i,EB} = \tilde{\theta}_{i,B}(\hat{\phi})$$

where $\hat{\phi}$ is the estimator of $\phi = (\beta, p, \sigma^2, \rho)$.

The mean square error for $\hat{\theta}_{i,EB}$ is

$$MSE\left(\hat{\theta}_{i,EB}\right) = E\left(\hat{\theta}_{i,EB} - \theta_i\right)^2 = E\left(\tilde{\theta}_{i,B}(\phi) - \theta_i\right)^2 + \ g_i(\phi) + o(n^{-1})$$

where $g_i(\phi) = O(n^{-1})$. That is,

$$MSE\left(\hat{\theta}_{i,EB}\right) = E\left[\tilde{g}_{i1}(y_i, \phi)\right] + \ g_i(\phi) + o(n^{-1}) = \ g_{i1}(\phi) + + g_i(\phi) + o(n^{-1})$$

and

$$\tilde{g}_{i1}(y_i, \phi) = E\left(\tilde{\theta}_{i,B}(\phi) - \theta_i\right)^2 = \pi_i\left(1; y_i, \theta^{(t)}\right)\frac{\sigma^2 \mathrm{D_i}}{\sigma^2 + \mathrm{D_i}} + \pi_i\left(0; y_i, \theta^{(t)}\right)\frac{\rho\sigma^2 \mathrm{D_i}}{\rho\sigma^2 + \mathrm{D_i}} +$$

$$\pi_i\left(1; y_i, \theta^{(t)}\right)\pi_i\left(0; y_i, \theta^{(t)}\right)(y_i - \ {x_i}^T\beta)^2\left[\frac{\sigma^2}{\sigma^2 + \mathrm{D_i}} - \frac{\rho\sigma^2}{\rho\sigma^2 + \mathrm{D_i}}\right]^2.$$

Note that, $g_{i1}(\phi)$ and $g_i(\phi)$ are unknown.

3.3 Bootstrapping method

To estimate $\hat{\Phi}$, we generate:

1. $\mathrm{z}_{\mathrm{i1}}^*: P(\mathrm{z}_{\mathrm{i1}}^* = 1) = \ \hat{p} = 1 - P(\mathrm{z}_{\mathrm{i1}}^* = 0)$;
2. $v_{i1}^*,\ v_{i2}^*$: $v_{i1}^* \sim N(0,\ \hat{\sigma}^2)$ and $v_{i2}^* \sim N(0,\ \hat{\rho}\hat{\sigma}^2)$, then $v_i^* = \mathrm{z}_{\mathrm{i1}}^* v_{i1}^* + (1 - \mathrm{z}_{\mathrm{i2}}^*)v_{i2}^*$ and $\theta_i^* = {x_i}^T\hat{\beta} + v_i^*$
3. $e_i^* \sim N(0,\ D_i)$ and compute $y_i^* = \ \theta_i^* + e_i^*$, and $\tilde{g}_{i1}\left(y_i^*, \hat{\phi}\right)$.

Based on y_i^* and θ_i^*, $i = 1, \ldots, n$, we obtain $\hat{\theta}_{i,EB}^*$. Repeat this procedure F times: $\{\theta_{i,f}^*, y_{i,f}^*, \hat{\theta}_{i,EB,f}^*, \hat{\phi}_f^*, f = 1, \ldots, F\}$.

$$M_{i1} = \frac{1}{F}\sum_{f=1}^{F}(\hat{\theta}_{i,EB,f}^* - \theta_{i,f}^*)^2, \quad \hat{g}_{i1,boot} = \frac{1}{F}\sum_{f=1}^{F}\tilde{g}_{i1}(y_{i,f}^*, \hat{\phi}_f^*).$$

$M_{i1} - \hat{g}_{i1,boot} + \tilde{g}_{i1}(y_i, \hat{\phi})$ is an approximate estimator of $MSE(\hat{\theta}_{i,EB}) = g_{i1}(\phi) + g_i(\phi) + o(n^{-1})$.

4. Application - Cash Rents Survey

4.1 Case study

The Cash Rents Survey is conducted on an annual basis by the United States Department of Agriculture's (USDA's) National Agricultural Statistics Service (NASS). The survey obtains cash rental rates from a representative sample of farmers and ranchers in the United States, excluding Alaska. This survey provides the basis for estimates of the current year's cash rents paid for irrigated cropland, non-irrigated cropland, and permanent pastureland. From the Cash Rents Survey, county, state, and national rental rates (dollars/acre) for each land-use category (irrigated, non-irrigated, and pasture) are published (see **Figure 1** for the 2020 county-level published cash rental rate estimates for irrigated land). Estimates of cash rental rates are useful to farmers, economists, and policy makers.

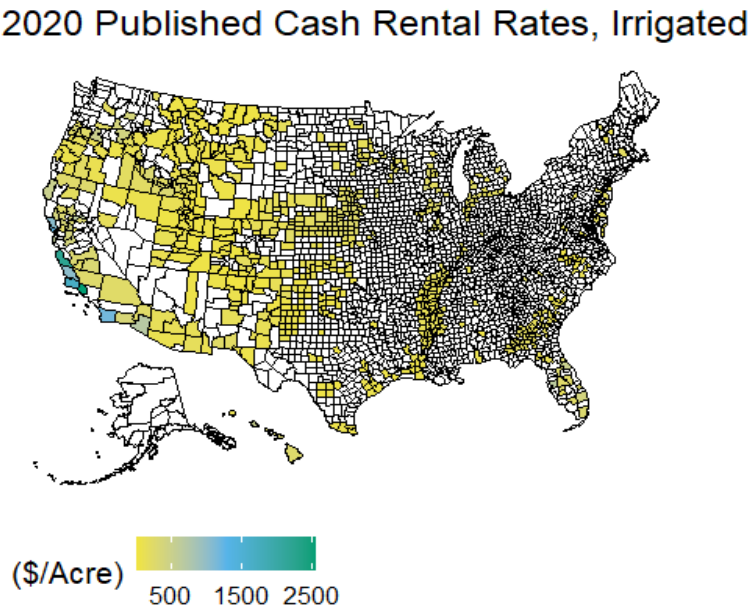

Figure 1: The 2020 county-level published cash rental rate official estimates for irrigated land

4.2 Sample design and design weights

The objective for Cash Rents Survey is to estimate rental acres, cash rents, and cash rental rates at the county level by land types: irrigated cropland, non-irrigated cropland, and pasture. The target population is the set of all U.S. (excluding Alaska) farms that rent land for cash during the reference year. The sampled population is a NASS-developed and maintained list frame of U.S. farms that rent land for cash. A stratified systematic sample design is drawn for the Cash Rents Survey. For a given state, we first stratify by county and next stratify by 10 general strata within each county. Then, we take a systematic sample from each stratum. Stratified sampling does not generate weight. It divides the full population into subpopulation. For a systematic sample, the sampling interval is the design weight for each unit (farm). The table below (see **Table 1** for stratified systematic sample design for Cash Rents Survey) is a systematic sample from 10 general strata for each county.

Stratified random sampling is a probability sampling technique in which the total population is divided into strata to complete the sampling process. Design weight for cash rents survey is from a systematic sample design. The weight is equal to the sample interval.

Table 1: Stratified systematic sample design for Cash Rents Survey

Strata	Description	Sampling Interval
98	Irrigated Cropland Acres > 10 % of total	1
96	Pasture Acres > 10% of total	1
94	Non-Irrigated Cropland Acres > 10% of total	1
92	Rent Paid for Land and Buildings > 10% of total	1
90	Irrigated Cropland Acres	1
80	Pasture Acres	2
70	Non-Irrigated Cropland Acres	2
60	Rent Paid for All Land and Buildings	3
50	Unknown Cash Rent Expenses	1
40	Land Rented or Leased from Others	40

Remark: when sampling interval is equal to 1, this stratum is a certainty stratum, that is, we select all farms in the stratum for survey.

4.3 Mixture model for cash rent survey

Realized sample sizes at the county level are often too small to support reliable direct estimates. In addition, outlier issues exist in some states. We conduct an exploratory data analysis in each state for different land types. It is important to check the distributions of survey estimates before fitting models. A density plot of the county-level 2020 California survey's direct estimates for irrigated land (see **Figure 2)** shows that the distribution is not unimodal but multimodal. A mixture of two components shows a good and parsimonious fit of the data based on the plot. Therefore, in this case study, we use the mixture modeling approach to accommodate the fit of multimodal distribution based on the model proposed in **Section 2**.

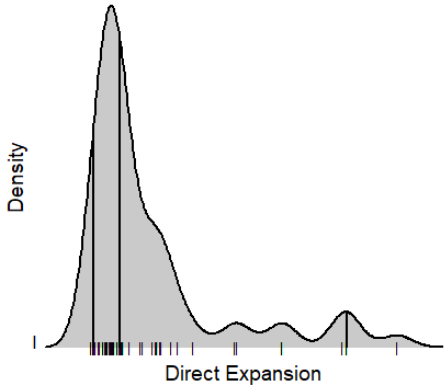

Figure 2: Density plot of the county-level 2020 California survey estimates for irrigated land

x_i are the covariates used in the model and include an intercept, the corresponding previous year county-level official estimates, the number of positive responses in the county, and the county-level National Commodity Crop Productivity Indices (NCCPIs). NCCPIs, which measure the quality of the soil for growing non-irrigated crops in climate conditions best suited for various crops.

To check the performance of the mixture model, we conduct the comparisons among mixture model (Mix), Fay-Herriot (FH) model, the published estimates, and the survey estimates. The Fay-Herriot model has the assumption that the distribution of the vi's is unimodal, not multimodal.

To evaluate the effectiveness of the estimator, we computed the following deviation measure for two model estimates and survey estimates from the NASS official statistics. The absolute relative deviation (ARD) is

$$ARD\ (\%) = \ 100 \times \frac{\left|y_i^{pub} - \hat{\theta}_i\right|}{y_i^{pub}},$$

where y_i^{pub} is the published estimates and $\hat{\theta}_i$ is the best empirical predictor from a model. External evaluation of potential models can shed light on their usefulness. Note that the published estimates would not be available for the current year. However, they are appropriate for use in assessing the quality and reasonableness of model-based estimates at the research stage.

The mixture model has a smaller mean and a smaller maximum ARD than the survey estimates and the FH model estimates, indicating that it more effectively accommodated the presence of outliers and the non-unimodal distribution (see **Table 2**). The maximum ARD represents the case for which an estimator is farthest from the published estimates. The maximum ARD for the mixture model estimates is almost 60% less than the survey estimates and 35% less than the FH model estimates. The minimum ARD represents the case for which the estimates are closest to the published estimates. The minimum ARD is smallest for the survey estimates. The FH estimates had the smallest median value of the ARD, and the mixture model had the smallest mean ARD.

We now turn to an examination of the performance of the MSE estimators associated with the different estimates (**Table 2)**. The MSE estimators for the mixture model are smaller than those for the survey and FH model estimates in all four summary measures.

Table 2: ARD (%) and MSE (%) summary measures based on survey, Fay-Herriot model, and mixture model

	ARD from Published Estimates (%)			**Mean Square Errors (%)**		
Model	**Survey**	**FH**	**Mix**	**Survey**	**FH**	**Mix**
Min	**0.02**	0.06	0.32	28.3	22.9	**21.7**
Median	12.5	**9.68**	10.7	4192.6	874.6	**805.1**
Mean	23.0	18.4	**17.9**	32539.8	10418.1	**9411.3**
Max	178.0	111.0	**72.1**	220107.1	93173.2	**69625.5**

5. Future Research

5.1 General mixture model

Based on a density plot of the county-level 2020 California survey's direct estimates for irrigated land (see **Figure 2**), the distribution of the vi's does not follow a normal distribution or a mixture of two normal distributions. A mixture of *K* normal distributions, where $K > 2$, provides a better fit for this common distribution.

Assume that the v_i's are independent and identically distributed random variables (iids), and the common distribution is a mixture of *K* normal distributions, with the *k*th component having mean *0* and variance σ_k^2, and the mixing proportion π_k, $k = 1,2,\cdots,K,\ 0 < \pi_k < 1,$ and $\sum_{k=1}^{K} \pi_k = 1$, where *K* is a known positive integer. We represent the distribution structure of v_i by

$$v_i = \sum_{k=1}^{K} z_{ik}\, v_{ik} = {z_i}^T v_i^*$$

where $z_i = (z_{i1}, z_{i2}, \cdots, z_{iK})^T \sim$ Multinomial $(1, \pi_1, \pi_2, \cdots, \pi_K)$, and $v_i^* = (v_{i1}, v_{i2}, \cdots, v_{iK})^T \sim MVN\left(0, Diag(\sigma_1^2, \sigma_2^2, \cdots, \sigma_K^2)\right)$. It is also assumed that z_i and v_i^* are independently distributed. Extending this work to the more general case would be useful.

5.2 Measurement Error Model

Ybarra and Lohr (2008) extend the area-level Fay-Herriot model to a measurement error model. The observed covariate is represented as the sum of a latent covariate and a mean zero measurement error. The measurement error in the covariate is uncorrelated with the sampling error in the response.

In the **Section 4**, one of covariates is the corresponding previous year county-level official estimates. So, consideration should be given to adding an additional equation: $W_i = x_i + u_i$, to the sampling equation and linking equation in eq. 1, where x_i is unobserved covariate and W_i is an estimate and observed. The resulting measurement error model is as follows:

$$\begin{cases} y_i = \theta_i + e_i & -\ sampling\ equation \\ \theta_i = x_i^T\beta + v_i & -\ linking\ equation \\ W_i = x_i + u_i & -\ measurement\ error \end{cases}$$

where the measurement error, u_i, is uncorrelated to model error, e_i. Their joint distribution is a multivariate normal distribution, that is,

$$\left(u_i^T, e_i\right)^T \sim MVN(0, \Psi_i)$$

where $\Psi_i = \begin{pmatrix} \Psi_{uui} & \Psi_{uei} \\ \Psi_{uei}^t & \Psi_{eei} \end{pmatrix}$.

The parameter of interest for the problem is $\theta_i = y_i - e_i$, and model parameters (β, σ^2) can be estimated though the empirical BLUP in the frequentist's framework, or the empirical Bayes estimator in the Bayesian framework.

Define observable quantity: $b_i = y_i - W_i^T\beta$. The best predictor of θ_i under squared error loss is

$$\tilde{\theta}_i = y_i - \hat{e}_i = y_i - E(e_i|b_i).$$

References

Datta, G.S. and Lahiri, P. (1995). Robust Hierarchical Bayes Estimation of Small Area Characteristics in the Presence of Covariates and Outliers, Journal of Multivariate Analysis, vol. 54, issue 2, 310-328.

Dempster, A.P., Laird, N. M., and Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data Via the EM Algorithm, Journal of the Royal Statistical Society: Series B (Methodological), Volume39, Issue1, pp. 1-22.

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(366a), 269–277.

Hall, P. and Maiti, T. (2006). On Parametric Bootstrap Methods for Small Area Prediction, Royal Statistical Society, Series B (Methodological), Vol. 68, No. 2, pp. 221-238.

Lahiri, P. and Rao, J. N. K. (1995). Robust Estimation of Mean Squared Error of Small Area Estimators, Journal of the American Statistical Association Vol. 90, No. 430, pp. 758-766.

Rao J.N.K. and Molina, I. (2015). Small area estimation, 2nd ed. New York, NY: John Wiley & Sons, Inc.

Ybarra, L. M. & Lohr, S. L. (2008). Small area estimation when auxiliary information is measured with error, Biometrika 95(4), 919–931.