Improved Estimation for June Area Survey Incorporating Several Information

Jae-Kwang Kim[1], Zhengyuan Zhu[1,2], and Shu Yang[1]

[1]Department of Statistics, Iowa State University, Ames, IA, USA
[2]Corresponding author, email: zhuz@iastate.edu

Abstract

In this paper we present a general methodology to improve the estimate of crop acreage by combining information from June Area Survey (JAS), administrative data from Farm Service Agency (FSA), and satellite imagery data summarized in Cropland Data Layer (CDL). Both a structural model and a measurement model are specified for the FSA and CDL data. The measurement error model is used to quantify the uncertainty in the estimate from each source, and the structural model is used to model the survey population. A parametric fractional imputation method is developed to estimate the parameters of the structural models, and a generalized method of moment is used to produce the final estimate of the crop acreage. The methodology is applied to produce improved estimate of crop acreage.

Keywords: Parametric fractional imputation; generalized method of moment; crop acreage; structural model

1 Introduction

Jun Area Survey is one of the largest annual NASS survey project which is designed to account for all land in all states except Alaska. All agricultural activities and land uses within the sampled segment boundaries are collected to provide direct estimates for acreage, cattle inventory, crop production summary, and many other publications. In this paper we focus on the estimation of acreage. Annually during the first two weeks of June nearly 11,000 segments of roughly the size of one square mile are selected for data collection. Though JAS can provide fairly reliable and timely estimates of acreage at the state level, those estimates can be improved by incorporating information from other sources such as the administrative record data from the Farm Service Agency (FSA) and the satellite imagery data summarized in Cropland Data Layer (CDL). If one is interested in small area estimation such as the county level acreage estimator, the use of all sources of information becomes critical as the sample size in JAS is not large enough to provide reliable estimates at county level. In this paper we consider the problem of combining three sources of information from JAS, FSA, and CDL to provide reliable, timely, and detailed estimation of acreage.

In JAS, the sample observations are obtained from a probability sampling. The FSA data is obtained from a voluntary participation of certain programs. The CDL data is obtained from the classification of satellite imagery data using a tree-based machine learning algorithm. To combine three source, we need county level estimates from each source. That is, we need

1. JAS estimate: $\hat{X}_i$ and $\hat{V}(\hat{X}_i)$
2. FSA estimate: $\hat{Y}_{1i}$ and $\hat{V}(\hat{Y}_{1i})$
3. CDL estimate: $\hat{Y}_{2i}$ and $\hat{V}(\hat{Y}_{2i})$

where subscript i denotes county i. In some case, we may have missing observation of $\hat{X}_i$ and $\hat{V}(\hat{X}_i)$ because no JAS sample is selected from that county. The counties with missing values are not used for parameter estimation but are considered for prediction. Let X_i be the ideal measurement of the acreage at the time of JAS operation which is free of sampling error. Our goal is to obtain a best prediction of X_i. A ratio-adjustment strategy can be used if estimation at a later reference time need to be considered.

From the FSA data, we can construct two models, one is the structural error model and the other is the measurement error models. The structural error model associated with the FSA data can be written as

$$Y_{1i} = \beta_0 + \beta_1 X_i + e_{1i}, \tag{1}$$

where Y_{1i} is the ideal measurement of the acreage at the time FSA data is collected, and $e_{1i} \sim (0, \sigma_1^2)$. The measurement error model is

$$\hat{Y}_{1i} = Y_{1i} + u_{1i}, \tag{2}$$

where $u_{1i} \sim (0, \hat{V}(\hat{Y}_{1i}))$. The structure model (1) is used to model the difference in acreage at the two reference time. The special case of $(\beta_0, \beta_1) = (0, 1)$ means that the (true) crop acres at the time of FSA estimation remain unchanged until JAS estimation, which it is not likely to hold. The model error e_{1i} represents the lack of fit when explaining the difference in the reference time.

Similarly, we can construct two models for CDL data. One is the structural error model, denoted by

$$Y_{2i} = \beta_0^* + \beta_1^* X_i + e_{2i},$$

where $e_{2i} \sim (0, \sigma_2^2)$, and the other is the measurement error model, denoted by

$$\hat{Y}_{2i} \quad = \quad Y_{2i} + u_{2i},$$

where $u_{2i} \sim (0, \hat{V}(\hat{Y}_{2i}))$.

Thus, for each source, we have two models. One is the structural error model and the other is the measurement error model. The structural error model is the model about the survey population while the measurement error model is the model about the estimates. We assume that the measurement variances can be estimated relatively accurately. In practice, we use a smoothing technique to reduce the variability associated with the variance estimates.

Note that The structural error model in (1) can be replaced by a parametric statistical model

$$Y_{1i} \sim f(y_{1i} \mid x_i; \theta_1),$$

where the parametric model $f(\cdot; \theta_1)$ is indexed by θ_1. For small crops, such parametric model approach can be useful.

In Section 2 we introduce the method for estimating the parameters in the structural error model. Prediction of X_i is covered in Section 3. We conclude with a discuss of possible refinement of the approach.

2 Parameter estimation

In this section, we first introduce the basic theory for parameter estimation of the measurement error models for FSA data. The parameter estimation for CDL data can be performed similarly. Note that only parameters in the structural error models need to be estimated. Using the theory of measurement error models (Fuller, 1987), a consistent estimator of (β_0, β_1) can be obtained by minimizing

$$Q^*(\beta_0,\beta_1) = \sum_{h=1}^{H} \begin{pmatrix} \bar{y}_{1h} - \beta_0 - \beta_1 \hat{\hat{X}}_h \\ \beta_1 (\bar{x}_h - \hat{\hat{X}}_h) \end{pmatrix}' \left\{ V \begin{pmatrix} \bar{y}_{1h} - \beta_0 - \beta_1 \hat{\hat{X}}_h \\ \beta_1 (\bar{x}_h - \hat{\hat{X}}_h) \end{pmatrix} \right\}^{-1} \begin{pmatrix} \bar{y}_{1h} - \beta_0 - \beta_1 \hat{\hat{X}}_h \\ \beta_1 (\bar{x}_h - \hat{\hat{X}}_h) \end{pmatrix}. \tag{3}$$

After some algebra, it can be shown that (3) reduces to

$$Q^*(\beta_0,\beta_1) = \sum_{h=1}^{H} \frac{(\bar{y}_{1h} - \beta_0 - \beta_1 \bar{x}_h)^2}{V(\bar{y}_{1h} - \beta_0 - \beta_1 \bar{x}_h)}. \tag{4}$$

As we can write

$$\bar{y}_{1h} - \beta_0 - \bar{x}_h \beta_1 = -a_h \beta_1 + b_h + \bar{e}_{1h},$$

we have

$$V(\bar{y}_{1h} - \beta_0 - \bar{x}_h \beta_1) = \sigma^2_{e,h} + (-\beta_1, 1)\,\Sigma_h\,(-\beta_1, 1)'. \tag{5}$$

where $\sigma^2_{e,h} = V(\bar{e}_{1h})$ and $\Sigma_h = V\{(a_h, b_h)'\}$. As we can obtain a consistent estimator of the variance-covariance matrix of (a_h, b_h), we can obtain $(\hat{\beta}_0, \hat{\beta}_1)$ minimizing $Q^*(\beta_0, \beta_1)$ in (4) if $\sigma^2_{e,h}$ is known. Thus, writing

$$Q^*(\beta_0,\beta_1) = \sum_{h=1}^{H} w_h(\beta_1)\,(\bar{y}_{1h} - \beta_0 - \beta_1 \bar{x}_h)^2, \tag{6}$$

where

$$w_h(\beta_1) = \left\{\sigma^2_{e,h} + (-\beta_1, 1)\,\Sigma_h\,(-\beta_1, 1)'\right\}^{-1},$$

we have

$$\frac{\partial}{\partial \beta_0} Q^* = 0 \quad \Longleftrightarrow \quad \sum_{h=1}^{H} w_h(\beta_1)\,(\bar{y}_{1h} - \beta_0 - \beta_1 \bar{x}_h) = 0$$

and so

$$\hat{\beta}_0 = \bar{y}_w - \hat{\beta}_1 \bar{x}_w, \tag{7}$$

where

$$(\bar{x}_w, \bar{y}_w) = \left\{ \sum_{h=1}^{H} w_h(\hat{\beta}_1) \right\}^{-1} \sum_{h=1}^{H} w_h(\hat{\beta}_1)\,(\bar{x}_h, \bar{y}_h).$$

Plugging (7) into (6), we have only to minimize

$$Q_1^*(\beta_1) = \sum_{h=1}^{H} w_h(\beta_1)\,\{\bar{y}_{1h} - \bar{y}_w - \beta_1(\bar{x}_h - \bar{x}_w)\}^2. \tag{8}$$

Thus, we need to find the solution to $\partial Q_1^*/\partial\beta_1 = 0$ where

$$\begin{aligned}\frac{\partial}{\partial\beta_1}Q_1^* &= \sum_{h=1}^{H}\left\{\frac{\partial}{\partial\beta_1}w_h(\beta_1)\right\}\{\bar{y}_{1h}-\bar{y}_w-\beta_1(\bar{x}_h-\bar{x}_w)\}^2 \\ &\quad -2\sum_{h=1}^{H}w_h(\beta_1)(\bar{x}_h-\bar{x}_w)\{\bar{y}_{1h}-\bar{y}_w-\beta_1(\bar{x}_h-\bar{x}_w)\}.\end{aligned}$$

Using

$$\frac{\partial}{\partial\beta_1}w_h(\beta_1) = -2\{w_h(\beta_1)\}^2\{\beta_1 V(a_h)-C(a_h,b_h)\},$$

and

$$\{\bar{y}_{1h}-\bar{y}_w-\beta_1(\bar{x}_h-\bar{x}_w)\}^2 \xrightarrow{p} \sigma_{e,h}^2+(-\beta_1,1)\Sigma_h(-\beta_1,1)' = 1/w_h(\beta_1),$$

where $\xrightarrow{p}$ denotes the convergence in probability, the solution to $\partial Q_1^*/\partial\beta_1 = 0$ satisfies

$$\hat{\beta}_1 = \frac{\sum_{h=1}^{H}w_h(\hat{\beta}_1)\{(\bar{x}_h-\bar{x}_w)(\bar{y}_{1h}-\bar{y}_{1w})-C(a_h,b_h)\}}{\sum_{h=1}^{H}w_h(\hat{\beta}_1)\left\{(\bar{x}_h-\bar{x}_w)^2-V(a_h)\right\}}. \tag{9}$$

Note that the weight $w_h(\beta_1)$ depends on β_1. Thus, the solution (9) can be obtained by an iterative algorithm. Once $\hat{\beta}_1$ is computed by (9), then $\hat{\beta}_0$ is obtained by (7).

Thus, the method-of-moment estimator of the parameters can be obtained by the following steps:

1. Obtain $\hat{\beta}_0^{(0)},\hat{\beta}_1^{(0)},\sigma_1^{(0)2}$ by the OLS.

2. For each t, compute

$$w_i^{(t)} = \left\{\hat{\sigma}_1^{(t)2}+\hat{V}(\hat{Y}_{1i})+\hat{\beta}_1^{(t)2}\hat{V}(\hat{X}_i)\right\}^{-1}$$

3. Compute

$$\hat{\beta}_1^{(t+1)} = \frac{\sum_i w_i^{(t)}\left(\hat{X}_i-\bar{X}_w^{(t)}\right)\left(\hat{Y}_{1i}-\bar{Y}_{1w}^{(t)}\right)}{\sum_i w_i^{(t)}\left\{\left(\hat{X}_i-\bar{X}_w^{(t)}\right)^2-\hat{V}(\hat{X}_i)\right\}}$$

$$\hat{\beta}_0^{(t+1)} = \bar{Y}_{1w}^{(t)}-\hat{\beta}_{1(t+1)}\bar{X}_w^{(t)}$$

and

$$\hat{\sigma}_1^{(t+1)2} = \frac{\sum_i w_i^{(t)}\left\{\left(\hat{Y}_{1i}-\hat{\beta}_0^{(t+1)}-\hat{\beta}_1^{(t+1)}\hat{X}_i\right)^2-\hat{V}(\hat{Y}_{1i})-\hat{\beta}_1^{(t+1)2}\hat{V}(\hat{X}_i)\right\}}{\sum_i w_i^{(t)}}$$

where

$$\left(\bar{X}_w^{(t)},\bar{Y}_{1w}^{(t)}\right) = \left(\sum_i w_i^{(t)}\right)^{-1}\sum_i w_i^{(t)}\left(\hat{X}_i,\hat{Y}_{1i}\right).$$

4. Goto Step 2 until convergence.

This is an iterative computation of method-of-moment estimators.

Instead of the MOM estimation, we can also consider the maximum likelihood estimation. In particular, we can use the parametric fractional imputation of Kim (2011) for maximum likelihood estimation. An iterative computation of the maximum likelihood estimator can be described as follows:

1. Obtain $\hat{\theta}_1^{(0)} = (\hat{\beta}_0^{(0)}, \hat{\beta}_1^{(0)}, \sigma_1^{(0)2})$ by the OLS.

2. Generate m values of X_i, denoted by $x_i^{*(1)}, \cdots, x_i^{*(m)}$, from $N(\hat{X}_i, \hat{V}(\hat{X}_i))$.

3. For each $x_i^{*(j)}$, assign fractional weights

$$w_{ij}^* \propto g(\hat{Y}_i \mid x_i^{*(j)}; \hat{\theta}_1^{(t)}, \hat{V}(\hat{Y}_i))$$

   where $g(\hat{Y}_i \mid x_i; \hat{\theta}_1^{(t)}, \hat{V}(\hat{Y}_i))$ is the density of $\hat{Y}_i$ conditional on x_i and $\sum_{j=1}^m w_{ij}^* = 1$. Because $\hat{Y}_{1i} = Y_{1i} + u_{1i}$, the density of $\hat{Y}_{1i}$ is the convolution of the density of Y_{1i} and the density of u_{1i}. If $u_{1i} \sim N(0, \hat{V}(\hat{Y}_{1i}))$ in (2), then

$$g(\hat{Y}_i \mid x_i; \hat{\theta}_1^{(t)}, \hat{V}(\hat{Y}_i)) = \int f(y \mid x_i^{*(j)}; \hat{\theta}_1^{(t)}) \times \phi\left\{ \frac{(\hat{Y}_i - y)}{(\hat{V}(\hat{Y}_i))^{1/2}} \right\} dy.$$

4. Using the fractionally imputed data in Step 3, solve the imputed score equation to obtain the updated parameters. That is, solve

$$\sum_i \sum_{j=1}^m E_{u_{1i}} \left\{ S(\theta_1; x_i^{*(j)}, \hat{Y}_{1i} - u_{1i}) \right\} = 0$$

   where $S(\theta_1; x, y) = \partial \log f(y \mid x; \theta)/\partial\theta_1$ is the score function and $E_{u_{1i}}(\cdot)$ is the expectation with respect to the distribution of $u_{1i} = \hat{Y}_{1i} - Y_{1i}$.

5. Using the updated parameters, check the convergence. Goto Step 3 until convergence.

This is a version of Monte Carlo EM (MCEM) algorithm using parametric fractional imputation. The resulting estimator is very close to the maximum likelihood estimator of θ_1 for sufficiently large m. Instead of the MCEM, one can also consider a Bayesian approach using Markov Chain Monte Carlo, which is computationally more challenging and also the convergence is hard to check.

3 Prediction

We now discuss best prediction from the measurement error models. The goal is to obtain an improved predictor of X_i. For the simple linear model setup, we can apply the GLS method

$$\begin{pmatrix} \hat{Y}_{1i} - \hat{\beta}_0 \\ \hat{Y}_{2i} - \hat{\beta}_0^* \\ \hat{X}_i \end{pmatrix} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_1^* \\ 1 \end{pmatrix} X_i + \begin{pmatrix} e_{1i} + u_{1i} \\ e_{2i} + u_{2i} \\ u_i \end{pmatrix}$$

where

$$\begin{pmatrix} e_{1i} + u_{1i} \\ e_{2i} + u_{2i} \\ u_i \end{pmatrix} \sim \left[\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \hat{V}(\hat{Y}_{1i}) + \hat{\sigma}_1^2 & 0 & 0 \\ 0 & \hat{V}(\hat{Y}_{2i}) + \hat{\sigma}_2^2 & 0 \\ 0 & 0 & \hat{V}(\hat{X}_i) \end{pmatrix} \right]$$

Thus, we can express

$$Y = Z\theta + e$$

where $e \sim (0, \hat{V})$ and obtain

$$\hat{\theta} = (Z'\hat{V}^{-1}Z)^{-1}Z'\hat{V}^{-1}Y. \tag{10}$$

The assumption of $Cov(e_{1i}, e_{2i} \mid X_i) = 0$ may not be true. However, accuracy of $\hat{V}$ is less critical.

In some counties, we may not have $\hat{X}_i$. (No JAS sample in the county). In this case, we can use

$$\begin{pmatrix} \hat{Y}_{1i} - \hat{\beta}_0 \\ \hat{Y}_{2i} - \hat{\beta}_0^* \end{pmatrix} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_1^* \end{pmatrix} X_i + \begin{pmatrix} e_{1i} + u_{1i} \\ e_{2i} + u_{2i} \end{pmatrix}$$

and apply the GLS estimator.

In the fractional imputation approach, the goal is to obtain a Monte Carlo approximation of the conditional expectation, $E(X_i \mid \hat{X}_i, \hat{Y}_{1i}, \hat{Y}_{2i})$. Since

$$f(X_i \mid \hat{X}_i, \hat{Y}_{1i}, \hat{Y}_{2i}) \propto f(X_i \mid \hat{X}_i) f(\hat{Y}_{1i}, \hat{Y}_{2i} \mid X_i),$$

and assuming $Cov(e_{1i}, e_{2i} \mid X_i) = 0$, the best predictor of X_i is obtained as a by-product of the computation. That is, we simply use

$$\hat{X}_i^* = \frac{\sum_{j=1}^m w_{1ij}^* w_{2ij}^* x_i^{*(j)}}{\sum_{j=1}^m w_{1ij}^* w_{2ij}^*},$$

where w_{1ij}^* are the fractional weights used for parameter estimation associated with the FSA data and w_{1ij}^* are the fractional weights used for parameter estimation associated with the CDL data.

Once $\hat{X}_i^*$ are obtained, the state-level estimate for state h is simply computed by

$$\hat{X}_h^* = \sum_{i \in U_h} \hat{X}_i^*, \tag{11}$$

where U_h is the set of counties in state h.

4 Discussion

Structural error model is important to link two different sources in the population. However, we observe only sample estimates, not the population values, and so measurement error model (or sampling error model) is also needed. The GLS method (or fractional imputation) is a useful tool for the parameter estimation and prediction. Variance estimation is not discussed and will be a topic of future research.

References

Fuller, W.A. (1987). *Measurement error models*, Wiley.

Kim, J.K. (2011). "Parametric fractional imputation for missing data analysis," *Biometrika*, **98**, 119–132.