District-level crop yield estimation using calibration approach

Kaustav Aditya*, Ankur Biswas, Ashok Kumar Gupta and Hukum Chandra

Indian Agricultural Statistics Research Institute, New Delhi 110 012, India

Estimation of major crop yield rates at the district level using calibration estimation technique is reported here when auxiliary information is available at the unit level only for the selected villages within each district and when the sampling design under consideration is two-stage equal probability without replacement. An estimator was developed for the complex sampling design under consideration using the calibration approach. Through evaluation using real data collected from a pilot survey, we found that the proposed calibration estimator performs better than the usual design-based Horvitz–Thompson estimator under two-stage sampling design.

Keywords: Calibration estimation technique, crop yield, two-stage sampling.

ONE of the most dependable methods to generate reliable estimates of the population parameters is sample survey. A typical survey objective is to estimate descriptive population parameters and analytical parameters, on the basis of a sample selected from a population of interest. The calibration estimation approach of Deville and Särndal[1], where the sampling weights are adjusted to make certain estimators match known population totals, is commonly used in survey sampling for increasing the efficiency of the design-based survey estimates. The generalized regression estimator is an example of a calibration estimator. Calibration consists of adjusting the weights such that estimates of the auxiliary variable(s) satisfy known totals (also referred to as control totals).

Deville and Särndal[1] estimated a finite population total in the presence of univariate or multivariate auxiliary information. Théberge[2] extended the calibration technique to estimate population parameters other than totals and means, and developed the technique when there is no solution to the calibration equation. He developed a method to compute a calibration estimator that used an arbitrary distance measure. For every distance measure there is a corresponding set of calibrated weights and a calibration estimator[1]. Calibration is also used to achieve consistency. The basic design-consistent Horvitz–Thompson[3] (HT) estimator is the most natural estimator to use if there is no auxiliary information available at the estimation stage. It weights data with the inverses of the inclusion probabilities of the sampled units. Such a weight is called a sampling weight. The properties of the HT estimator can be improved by following calibration estimation technique when auxiliary information is available. This procedure adjusts the sampling weights by multipliers known as calibration factors that make the estimates agree with known totals. The resulting weights are called calibration weights and the resulting estimates will be design-consistent with smaller sampling variance than the HT estimator.

Suppose we are interested in computing the total value of variable Y. Let us assume that the whole population $U = \{1,\ldots, k,\ldots, N\}$ consists of N elements. From this population we draw at random sample s of size n without replacement. Let π_i denote the first-order inclusion probability, i.e. $\pi_i = p_r\ (i \in s)$ and $d_i = (1/\pi_i)$, the sampling weight defined as the inverse of the inclusion probability for unit i. Let $\pi_{ij} = p_r\ (i \text{ and } j \in s)$. Our objective is to estimate the population total of variable y which is given as

$$Y = \sum_{i=1}^{N} y_i. \tag{1}$$

Classical estimator of the population total (eq. (1)) is the HT estimator, which is given by the following formula

$$\hat{Y}_{\mathrm{HT}} = \sum_{i=1}^{n} \frac{y_i}{\pi_i} = \sum_{i=1}^{n} d_i y_i. \tag{2}$$

Its variance under the sampling design is given as

$$V(\hat{Y}_{\mathrm{HT}}) = \sum_{i=1}^{N}\sum_{j=1}^{N} (\pi_{ij} - \pi_i\pi_j)\frac{y_i}{\pi_i}\frac{y_j}{\pi_j}. \tag{3}$$

Now let us suppose that values of the auxiliary variable x, i.e. $\{x_i, i = 1,\ldots, N\}$ are available and $X = \sum_{i=1}^{N} x_i$, the population total is known.

Ideally we would like $\sum_{i=1}^{n} d_i x_i = X$. Sometimes this is not true. The idea behind calibration estimators is to find weights w_i, $i = 1,\ldots, n$, close to d_i, based on a distance function, such that $\sum_{i=1}^{n} w_i x_i = X$.

A simple case considered by Deville and Särndal[1] is the minimization of chi-square-type distance function given by

$$\sum_{i=1}^{n} \frac{(w_i - d_i)^2}{d_i q_i},$$

subject to the constraint equation $\sum_{i=1}^{n} w_i x_i = X$; where q_i are suitably chosen weights called tuning parameter.

The weight w_i is then obtained as

$$w_i = d_i + \frac{d_i q_i x_i}{\sum_{i=1}^{n} d_i q_i x_i^2}\left(X - \sum_{i=1}^{n} d_i x_i\right).$$

*For correspondence. (e-mail: katu4493@gmail.com)

Substituting the value of w_i in calibrator estimator $\hat{Y}_c = \sum_{i=1}^{n} w_i y_i$, only gives

$$\hat{Y}_c = \sum_{i=1}^{n} d_i y_i + \frac{\sum_{i=1}^{n} d_i q_i x_i y_i}{\sum_{i=1}^{n} d_i q_i x_i^2} \left(X - \sum_{i=1}^{n} d_i x_i \right)$$
$$= \hat{Y}_{HT} + \hat{B}(X - \hat{X}_{HT}),$$

where

$$\hat{B} = \frac{\sum_{i=1}^{n} d_i q_i x_i y_i}{\sum_{i=1}^{n} d_i q_i x_i^2}.$$

Written in this form, we see that $\hat{Y}_c$ is same as the linear GREG estimator[4].

In fact, the GREG estimator is a special case of the calibration estimator when the chosen distance function is the chi-square distance[1]. In the GREG approach the predicted values are generated using an assisting model, whereas in calibration approach they do not depend on any assumption about the assisting model, an assumed relationship (linear, nonlinear, generalized linear, mixed, and so on) between study variable and auxiliary variable[5].

The variance of the calibration estimator is given as

$$V(\hat{Y}_c) = V(\hat{Y}_{HT} + \hat{B}(X - \hat{X}_{HT}))$$
$$= \sum_{i=1}^{N} \sum_{j=1}^{N} \Delta_{ij} (d_i (y_i - Bx_i))(d_j (y_j - Bx_j)).$$

The estimator of variance of the proposed calibration estimator is given as

$$\hat{V}(\hat{Y}_c) = \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{\Delta_{ij}}{\pi_{ij}} (w_i e_i)(w_j e_j),$$

where $e_i = y_i - \hat{\beta} x_i$ and $\Delta_{ij} = (\pi_i \pi_j - \pi_{ij})$.

Calibration estimators of unistage sampling designs have been proposed by many researchers[5–13], but no calibration estimator was developed for the case of multistage sampling designs which are generally used for large- to medium-scale surveys. Aditya *et al.*[14] developed calibration estimators of the population total under multistage sampling design assuming positive correlation between the study variable and auxiliary variable, and availability of auxiliary information at cluster level (i.e. all primary stage units; psus). They developed higher-order calibration estimators for precise estimation of variance under two-stage sampling design[14]. Here we report the calibration estimator when auxiliary information is available at the secondary stage unit (ssu) level only for the selected psus.

We have considered the simple case where information on only one auxiliary variable is available. Let the population of elements $U = \{1,\ldots, k,\ldots, N_I\}$ be partitioned into clusters (i.e. psus), $U_1, U_2,\ldots, U_i,\ldots, U_{N_I}$. The size of U_i is denoted as N_i. We have

$$U = \bigcup_{i=1}^{N_I} U_i \quad \text{and} \quad N = \sum_{i=1}^{N_I} N_i.$$

At stage one, a sample of psus, s_I, of size n_I is selected from U_I according to the design $p_I(.)$ with the inclusion probabilities π_{Ii} and π_{Iij} at the psu level. Given that the psu U_i is selected at the first stage, a sample s_i of size n_i units is drawn from U_i according to some specified design $p_i(.)$ with inclusion probabilities $\pi_{k/i}$ and $\pi_{kl/i}$. For the second stage sampling we assume the invariance and independence property. The whole sample of elements and its size are then respectively

$$s = \bigcup_{i=1}^{S_I} s_i \quad \text{and} \quad n_s = \sum_{i=1}^{n_I} n_i.$$

The inclusion probabilities at the first stage are given as

$$\pi_{Ii} = \Pr(i \in s_I),$$

$$\pi_{Iij} = \begin{cases} \Pr(i \text{ and } j \in s_I), & i \text{ and } j \text{ are different psus,} \\ \pi_{Ii}, & i \text{ and } j \text{ are the same psus.} \end{cases}$$

The inclusion probabilities for the second stage are given as

$$\pi_{k/i} = \Pr(k \in s_i \mid i \in s_I) \text{ and,}$$

$$\pi_{kl/i} = \begin{cases} \Pr(k \text{ and } l \in s_i \mid i \in s_I), & k \text{ and } l \text{ are different,} \\ \pi_{k/i}, & k \text{ and } l \text{ are the same.} \end{cases}$$

Let the study variable be y_k, which is observed for $k \in s$. The parameter to estimate is the population mean

$$\bar{Y} = \frac{1}{N} \sum_{i=1}^{N} y_k = \frac{1}{N} \sum_{i=1}^{N_I} \bar{y}_{i.},$$

where

$$\bar{y}_{i.} = \frac{1}{N_i} \sum_{k=1}^{N_i} y_k = i\text{-th psu total.}$$

In the present situation, auxiliary information is available for ssu for the selected psus only, i.e. the auxiliary information x_k is known for all elements $k \in s$, while the correct value of $(1/N_i)\sum_{k=1}^{N_i} x_k$ is available for each sampled psu, and correlation between the study variable and auxiliary variable is assumed to be positive.

The simple HT estimator of the population mean in this case will be

$$\bar{y}_{\mathrm{HT}} = \sum_{i=1}^{n_I} a_{Ii} \sum_{k=1}^{n_i} a_{k/i}\, y_k .$$

The proposed calibration estimator of the population mean in this case is given as

$$\bar{y}_{\mathrm{cal}} = \sum_{i=1}^{n_I} a_{Ii} \sum_{k=1}^{n_i} w_k^* y_k , \tag{5}$$

where w_k^* is the calibrated weight corresponding to the design weight $a_{k/i}$. Here, we minimize the chi-square-type distance function using Lagrangian multiplier technique as described in the earlier cases and obtain the calibrated weight. We minimize

$$\sum_{k=1}^{n_i} \frac{(w_k^* - a_{k/i})^2}{a_{k/i} q_k^*}, \text{ such that } \sum_{k=1}^{n_i} w_k^* x_k = \sum_{k=1}^{N_i} x_k .$$

Hence the calibrated weight is obtained as

$$w_k^* = a_{k/i} + \frac{a_{k/i} q_k^* x_k}{\sum_{k=1}^{n_s} a_{k/i} q_k^* x^2{}_k} \left(\sum_{k=1}^{N_i} x_k - \sum_{k=1}^{n_i} a_{k/i} x_k \right).$$

After considering $q_k^* = 1,$ the estimator becomes

$$\bar{y}_{\mathrm{cal}} = \sum_{i=1}^{n_I} a_{Ii} \left[\sum_{k=1}^{n_i} a_{k/i}\, y_k + \sum_{k=1}^{n_i} \frac{a_{k/i} x_k y_k}{\sum_{k=1}^{n_s} a_{k/i} x^2{}_k} \times \left(\sum_{k=1}^{N_i} x_k - \sum_{k=1}^{n_i} a_{k/i} x_k \right) \right]$$

$$= \sum_{i=1}^{n_I} a_{Ii} \left[\sum_{k=1}^{n_i} a_{k/i}\, g_{ks} y_k \right].$$

Now considering $q_k^* = (1/z_i)$ gives

$$\bar{y}_{\mathrm{cal}} = \sum_{i=1}^{n_I} a_{Ii} \left[\frac{\sum_{k=1}^{n_i} a_{k/i} y_k \left(\sum_{k=1}^{N_i} x_k \right)}{\sum_{k=1}^{n_i} a_{k/i} x_k} \right].$$

The above estimator takes the form of a ratio estimator under this condition. Under an equal probability without replacement design (SRSWOR) at both the stages, it reduces to

$$\bar{y}_{\mathrm{cal}} = \frac{1}{n_I} \sum_{i=1}^{n_I} \frac{\left(\frac{1}{n_i} \sum_{k=1}^{n_i} y_k \right) \left(\frac{1}{N_i} \sum_{i=1}^{N_i} x_k \right)}{\left(\frac{1}{n_i} \sum_{k=1}^{n_i} x_k \right)} = \frac{1}{n_I} \sum_{i=1}^{n_I} \frac{(\bar{y}_{i.})(\bar{X}_{i.})}{(\bar{x}_{i.})}.$$

It can be seen that the calibration estimator takes the form of a simple ratio estimator under two-stage sampling design[15,16]. The approximate variance of the proposed estimator is obtained by first-order Taylor series linearization technique[17] as

$$V(\bar{y}_{\mathrm{cal}}) = \sum_{i=1}^{N_I} \sum_{j=1}^{N_I} \Delta_{Iij} \frac{t_{y_i}}{\pi_{Ii}} \frac{t_{y_j}}{\pi_{Ij}} + \sum_{i=1}^{N_I} \frac{1}{\pi_{Ii}} \sum_{k=1}^{N_i} \sum_{l=1}^{N_i} \Delta_{kl/i} \frac{E_k''}{\pi_{k/i}} \frac{E_l''}{\pi_{l/i}}, \tag{6}$$

where

$$E_k'' = y_k - \beta'' x_k, \; t_{y_i} = \sum_{k=1}^{N_i} y_k, \; \Delta_{Iij} = (\pi_{Iij} - \pi_{Ii}\pi_{Ij}),$$

$$\Delta_{kl/i} = \pi_{kl/i} - \pi_{k/i}\pi_{l/i} \text{ and } \beta'' = \frac{\sum_{k=1}^{N} y_k x_k}{\sum_{i=1}^{N} x_k^2}.$$

The approximate form of the estimator of variance of the calibration estimator is

$$\hat{V}(\bar{y}_{\mathrm{cal}}) = \frac{1}{2} \sum_{i=1}^{n_I} \sum_{j=1}^{n_I} d_{Iij} \left(\frac{\hat{t}_{y_{i\pi}}}{\pi_{Ii}} - \frac{\hat{t}_{y_{j\pi}}}{\pi_{Ij}} \right)^2 + \frac{1}{2} \sum_{j=1}^{n_I} \frac{1}{\pi_{Ii}^2} \sum_{k=1}^{n_i} \sum_{l=1}^{n_i} d_{kl/i} (w_k^* e_{ks} - w_l^* e_{ls})^2, \tag{7}$$

Table 1. Correlation between crop yield (study variable) and seed rate (auxiliary variable) for paddy and wheat

District	Paddy	Wheat
Barabanki	0.31	0.13
Bareilly	0.18	0.14
Bulandshahar	0.07	0.09

Table 2. Observed values in three districts for *kharif* paddy and *rabi* wheat for the agricultural year 2013–14

	Paddy			Wheat		
District	Average area (ha)	Production (kg)	Average seed rate (kg/ha)	Average area (ha)	Production (kg)	Average seed rate (kg/ha)
Barabanki	0.88	1,059,107	Barabanki	0.88	1,059,107	Barabanki
Bareilly	0.97	893,828	Bareilly	0.97	893,828	Bareilly
Bulandshahar	0.98	794,642	Bulandshahar	0.98	794,642	Bulandshahar

Table 3. Comparison of calibration estimator with Horvitz–Thomson estimator for yield (kg/ha)

	Paddy				Wheat			
District	$\bar{y}_{\mathrm{HT}}$	SE	$\bar{y}_{\mathrm{cal}}$	SE	$\bar{y}_{\mathrm{HT}}$	SE	$\bar{y}_{\mathrm{cal}}$	SE
Barabanki	3853.39	1.31	3638.87	0.80	3791.89	3.23	5136.72	0.43
Bareilly	3128.14	3.24	3447.18	0.45	4158.57	3.42	4117.82	0.47
Bulandshahar	4037.01	1.08	4701.51	0.79	4686.76	2.49	3601.40	0.90

where

$$e_{ks} = y_k - \hat{\beta}'' x_k, \quad d_{Iij} = \frac{(\pi_{Ii}\pi_{Ij} - \pi_{Iij})}{\pi_{Iij}},$$

$$d_{kl/i} = \frac{(\pi_{k/i}\pi_{l/i} - \pi_{kl/i})}{\pi_{kl/i}} \text{ and } \hat{\beta}'' = \frac{\sum_{i=1}^{n_s} a_k y_k x_k}{\sum_{i=1}^{n_s} a_k x^2{}_k}.$$

Data on crop yield, crop area and seed rate of paddy in *kharif* and wheat in *rabi* season for the agricultural year 2013–14 collected from selected farmers in three districts of Uttar Pradesh (UP), viz. Barabanki, Bareilly and Bulandshahar were used for the empirical study. This is a part of the data collected from the pilot survey carried out for estimation of seed, feed and wastage ratios of the major food grain crops by the Division of Sample Survey of ICAR-Indian Agricultural Statistics Research Institute, New Delhi. The sampling design adopted for data collection is that of stratified two-stage random sampling with district as stratum, village as psu and farmers growing food-grain crops and having livestock as ssu. In a district, a sample of 20 villages was allocated to different tehsils/blocks in proportion to the number of villages and selected by simple random sampling without replacement (SRSWOR): If any selected village was found to be inhabited, it was substituted with another village falling in the same tehsil/block.

Table 1 provides the observed values of total production (kg) in the three districts for rice in *kharif* and wheat in *rabi* season. The correlation between crop yield and yield rate was found to be positive in all the six cases (Table 2). However, no assumption about the assisting model between the study variable and auxiliary variable was made for the calibration approach. Necessary computations required for the estimation were made using the R software.

The value of the population-level seed rate was kept fixed throughout the empirical study as 25 kg/ha for rice and 100 kg/ha for wheat in all the three districts.

For each crop, estimates of calibration ($\bar{y}_{\mathrm{cal}}$) and HT ($\bar{y}_{\mathrm{HT}}$) estimators were generated from the given data. Coefficient of variation (CV) of both the estimators was also worked out. Table 3 provides results obtained from the above calculations.

A close perusal of Table 3 shows that the estimate of paddy yield using HT estimator lies between 3127.39 and 4037.01 kg/ha at the district level along with the %CV varying between 1.08 and maximum 3.24, whereas the yield estimates generated using the proposed calibration ratio-type estimator lie between 3638.87 and 4701.51 kg/ha with %CV varying between 0.45–0.80 at

the district level. There was significant improvement in the estimators by the use of auxiliary information through calibration estimation technique for estimation of rice yield. In Bareilly district, there was maximum improvement in %CV of the calibration ratio-type estimator ($\bar{y}_{\text{cal}}$) over the simple HT estimator ($\bar{y}_{\text{HT}}$) for estimation of paddy yield when the sampling design under consideration was two-stage equal probability without replacement sampling design at each stage of selection. Further, the yield estimates of wheat crop varied from 3797.89 to 4686.76 kg/ha in case of HT estimator, whereas it varied from 3601.40 to 5136.72 kg/ha in case of the proposed calibration ratio-type estimator. The %CV varied from 2.49 to 3.42 in case of the HT estimator and it varied from 0.43 to 0.90 in case of the proposed calibration ratio-type estimator. So for estimation of wheat yield in the above-mentioned districts of UP, it can be seen that calibration ratio-type estimator of crop yield performs better that the usual HT estimator with respect to improvement in %CV under two-stage equal probability without replacement sampling design.

It can be concluded that for estimation of crop yield, the proposed estimator is more efficient than the HT estimator with respect to %CV under two-stage equal probability without replacement sampling design. Further, it can be concluded that no prior assumptions are made about the assisting model for formation of estimators with the help of auxillary informations, calibration estimation technique can be treated as a better alternative.

1. Deville, J. C. and Särndal, C. E., Calibration estimators in survey sampling. *J. Am. Stat. Assoc.*, 1992, **87**, 376–382.
2. Théberge, A., Extension of calibration estimators in survey sampling. *J. Am. Stat. Assoc.*, 1999, **94**, 635–644; http://dx.doi.org/10.1080/01621459.1999.10474157.
3. Horvitz, D. G. and Thompson, D. J., A generalization of sampling without replacement from a finite universe. *J. Am. Stat. Assoc.*, 1952, **47**, 663–685.
4. Cassel, C. M., Särndal, C. E. and Wretman, J. H., Some results on generalized difference estimation and generalized regression estimation for finite population. *Biometrica*, 1976, **63**, 615–620.
5. Singh, S., Survey statisticians celebrate Golden Jubilee Year – 2003 of the linear regression estimator. *Metrika*, 2006, **63**, 1–18.
6. Estevao, V. M. and Särndal, C. E., A new perspective on calibration estimators. In Joint Statistical Meeting – Section on Survey Research Methods, 2003, pp. 1346–1356.
7. Farrell, P. J. and Singh, S., Penalized chi square distancefunction in survey sampling. In Proceedings of Joint Statistical Meeting, New York (Available on CD), 2002, vol. 15.
8. Farrell, P. J. and Singh, S., Model-assisted higher order calibration of estimators of variance. *Aust. New Zealand J. Stat.*, 2005, **47**(3), 375–383.
9. Kott, P. S., An overview of calibration weighting. *Joint Statistical Meeting – Section of Survey Methods*, 2003, pp. 2241–2252.
10. Montanari, G. E. and Ranalli, G., Nonparametric model calibration estimation in survey sampling. *J. Am. Stat. Assoc.*, 2005, **100**(472), 1429–1442.
11. Singh, S., *Advanced Sampling Theory with Applications: How Michael 'Selected' Amy*, Kluwer Academic Publisher, The Netherlands, 2003, vols 1 & 2, pp. 1–1247.
12. Singh, S., Golden and Silver Jubilee Year – 2003 of the linear regression estimators. Presented at the Joint Statistical Meeting, Toronto (available on the CD), 2004, pp. 4382–4389.
13. Wu, C. and Sitter, R. R., A model-calibration approach to using complete auxiliary information from survey data. *J. Am. Stat. Assoc.*, 2001, **96**, 185–193.
14. Aditya, K., Sud, U. C., Chandra, H. and Biswas, A., Calibration based regression type estimator of the population total under two stage sampling design. *J. Indian Soc. Agric. Stat.*, 2016, **70**(1), 19–24.
15. Aditya, K. and Sud, U. C., Higher order calibration estimators under two stage sampling. In *Statistics and Informatics in Agricultural Research*, Excel India Publication, New Delhi, 2015.
16. Cochran, W. G., *Sampling Techniques*, John Wiley & Sons, New York, 1977, 3rd edn.
17. Särndal, C. E., Swensson, B. and Wretman, J., *Model-Assisted Survey Sampling*, Springer-Verlag, New York, 1992.

Received 27 May 2016; revised accepted 22 December 2016

doi: 10.18520/cs/v112/i09/1927-1931

Estimation of carrying capacity of livestock farm based on maximum phosphorus load of farmland and GIS spatial analysis technology

Bojie Yan[1], Jingjie Yan[2] and Wenjiao Shi[3,*]

[1]Department of Geography, Minjiang University, Fuzhou, 350108, China
[2]College of Telecommunications and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210003, China
[3]Key Laboratory of Land Surface Pattern and Simulation, Institute of Geographic Sciences and Natural Resources Research, CAS, Beijing 100101, China

To avoid the environmental pollution caused by livestock manure and provide rational layout of livestock farm, we estimated the livestock manure phosphorus load by the excretion coefficient method and have developed a livestock manure nutrient distribution model. The livestock manure phosphorus was distributed to farmlands using this model and spatial analysis technology. The carrying capacity of livestock farms was calculated based on the maximum livestock manure phosphorus carrying capacity of farmlands and expressed in pig for the Shangjie town, China. The results showed that the maximum, minimum, average and total livestock manure phosphorus carrying capacity of farmlands was about 55.97, 0.74, 12.21 and 13,382.90 kg respectively, and the total load of

*For correspondence. (e-mail: shiwj@lreis.ac.cn)